REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
July 25 2002

3. REPORT TYPE AND DATES COVERED
Final report 14 June 2001 - 13 June 2002

4. TITLE AND SUBTITLE
Design, development and testing of vision algorithms for the detection of human shapes.

5. FUNDING NUMBERS
C-N68171-01-M-5875

R&D 9095-AN-015

6. AUTHOR(S)
Massimo Bertozzi, Alberto Broggi, Alessandra Fascioli, Paolo Lombardi, and Amos Tibaldi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Dip. Informatica e Sistemistica
Università di Pavia
Via Ferrata, 1
27100 Pavia
ITALY

8. PERFORMING ORGANIZATION REPORT NUMBER
0003

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Naval Regional Contracting Center, Det. London
Government Buildings, Block 2, Wing 12
Lime Grove, Ruislip, Middlesex HA4 8BX,
UNITED KINGDOM

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

This report presents the research work developed under contract number N68171-01-M-5857 with the aim of localizing human shapes in day-light images.
First an outlook of the problem is given, along with the description of the chosen approach and motivation, then the schedule of the activities performed during the research period is presented.
The main part of this report is centered on the description of the algorithm developed.
A final section discusses the results obtained in different conditions and scenarios, giving the qualitative and quantitative performance achieved on sample sequence of images.
The publications resulting from this research are also included.

14. SUBJECT TERMS
Human shape detection, vision algorithms.

15. NUMBER OF PAGES
12

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT

AD

Design, development and testing of vision algorithms for the detection of human shapes.

Final Technical Report
by

Massimo Bertozzi, Alberto Broggi, Alessandra Fascioli, Paolo Lombardi, and Amos Tibaldi
June 2001 - June 2002

United States Army

EUROPEAN RESEARCH OFFICE OF THE U.S. ARMY

London, England

CONTRACT NUMBER N68171-01-M-5857

UNIVERSITÀ DI PAVIA

Pavia, Italy

Approved for Public Release; distribution unlimited

Abstract

This report presents the research work developed under contract number N68171-01-M-5857 with the aim of localizing human shapes in day-light images. First an outlook of the problem is given, along with the description of the chosen approach and motivation, then the schedule of the activities performed during the research period is presented. The main part of this report is centered on the description of the algorithm developed. A final section discusses the results obtained in different conditions and scenarios, giving the qualitative and quantitative performance achieved on sample sequence of images. The publications resulting from this research are also included.

Keywords: Human shape detection, Vision algorithms.

Contents

1 Description of the problem

The research activity developed under this contract was aimed at the design, development, and test of vision algorithms for the detection of human shapes. The robust localization of human shapes in unstructured environments (e.g. a battlefield scenario) is a problem of basic importance for military applications.

Many systems have been developed based on different sensors, but vision is the only means of sensing the environment that is not based on the emission of signals. In fact, radars or lasers are based on the measurement of the alterations of signals emitted by the sensors themselves. The possibility of perceiving the environment in a passive way can be strategical in this field of application. Moreover the description of the scene provided by visual sensors is extremely rich and contains many details not perceivable by other sensors.

The research group of the Universities of Parma and Pavia have been active in the field of Intelligent Transportation Systems for years. Their main research field is artificial vision for intelligent vehicles. An outstanding result was the development a prototype vehicle within the ARGO Project. This project is aimed at developing and testing innovative solutions to be included on future vehicles to increment safety and reduce accidents. The first main result of the project was a system able to warn the driver in dangerous conditions. After the successful demonstration of this device, the research continued, mainly focusing on automating some of the driving tasks. In 1998 a vehicle prototype was demonstrated to the public and to the scientific community thanks to a tour through Italy (2000+ km) driven in automatic mode. The vehicle is equipped with low-cost cameras and a standard PC. It can compute the geometry of the road, localize obstacles and pedestrians, and track the vehicle ahead in real-time; the result of the processing is used to drive a motor on the steering wheel that permits to keep the vehicle within the driving lane, overtake slower vehicles, and follow the vehicle ahead.

2 Description of the activities performed

The activities carried out during the contract period were:

- A large-spectrum analysis of the state of the art in the field of pedestrian detection was developed. For each considered project both the algorithms and the hardware platforms were analyzed, and a classification of the different approaches was done.
- The working environment was defined as having the following characteristics.
  - Only flat scenarios were considered (such as road environments).
  - Friendly environments were considered (simple and uncluttered scenes with a small number of un-occluded pedestrians).
  - Images were acquired during day time. Critical and extreme situations such as direct sunlight, reflections or strong shadows were not addressed.
- The system requirements were identified.
  - Standard cameras (in the visible spectrum) were used.
  - Since also pedestrians standing still should be detected, no motion cues can be exploited and thus only single shots were considered (no feature tracking among image sequences).
  - Among the many possible poses of pedestrians, the algorithm will have to work on standing and walking pedestrians.
  - Pedestrians will have to be non overlapping and non occluded. They should have a sufficient contrast with the background to allow detection in images in the visible spectrum.
  - Since the detection will be performed on digital images, pedestrians should be represented by a sufficiently large number of pixels. This requirement generates a relation between camera aperture, image resolution, pedestrian size and distance. As an example, assuming a target height range of 50-100 pixels, a camera aperture of 40 degrees, an image resolution of 384 x 288 pixels, and a typical pedestrian height of 1.7 meters, the pedestrian distance will range from 10 to 25 meters.

- The specification of the computing system were defined and two PCs were purchased. The two PCs are based on an AMD processor at 1.3 GHz. Moreover, a VCR and a video acquisition board were purchased. The board is able to acquire pairs of stereo images in real-time directly on the PC's RAM.
- Preliminary image acquisition tests were conducted in different scenarios.
- The most salient features characterizing the human shape in visible spectrum images were selected on the basis of the already acquired image sequences. The main invariants of human shape that were devised are:
  - strong vertical symmetry
  - high density of vertical edges
  - aspect ratio varying within a specific range
  - size varying within a specific range
  - shape characteristics such as head and legs.
- The basic functions of the algorithm were defined and a preliminary implementation was developed. No real-time requirements were set upon this first implementation, as problems strictly connected to the algorithm were addressed first. Optimization of the algorithm allowing fast processing is set as a long-term goal.
- Preliminary results were evaluated and a critical analysis on the system behavior was performed.
- A further development of the algorithm has been carried out, focusing in particular on the medium-level candidates validation phase.
- A working environment was realized, featuring easy-to-access windows for the display of the images under analysis. This environment includes facilities for development of new functionalities and tuning of the parameters. Moreover it allows debugging and easy refinement of the algorithm.
- Further enhancements of the algorithm were accomplished. More specifically a number of filters were implemented to eliminate false positives such as poles and trees, to discard symmetrical areas between two human shapes and to refine the candidates.
- Testing of the system on newly acquired images in environments both similar and different from the one of the original test sequences were performed.
- The performance of the system was evaluated, in the form of the percentage of correctly detected human shapes and the number of false positive detections.

3 Description of the algorithm developed

The algorithm developed is based on the search for specific morphological characteristics of the human shape. No motion cues are currently used, thus the algorithm is not limited to the detection of moving people. Anyway, due to the extremely large variety of poses, texture, color, and clothing, the detection of people is a very challenging problem. In order to simplify this task, some assumptions were necessary: the current algorithm has been designed to localize standing people only, assuming a sufficient contrast with the background. The basics of this functionality are the search for symmetries and specific perspective and aspect ratio constraints.

Initially, attentive vision techniques relying on the search for specific characteristics of pedestrians such as vertical symmetry and strong presence of edges, allow to select interesting regions likely to contain pedestrians. Then, such candidate areas are validated verifying the actual presence of pedestrians by means of a shape matching technique based on the application of autonomous agents.

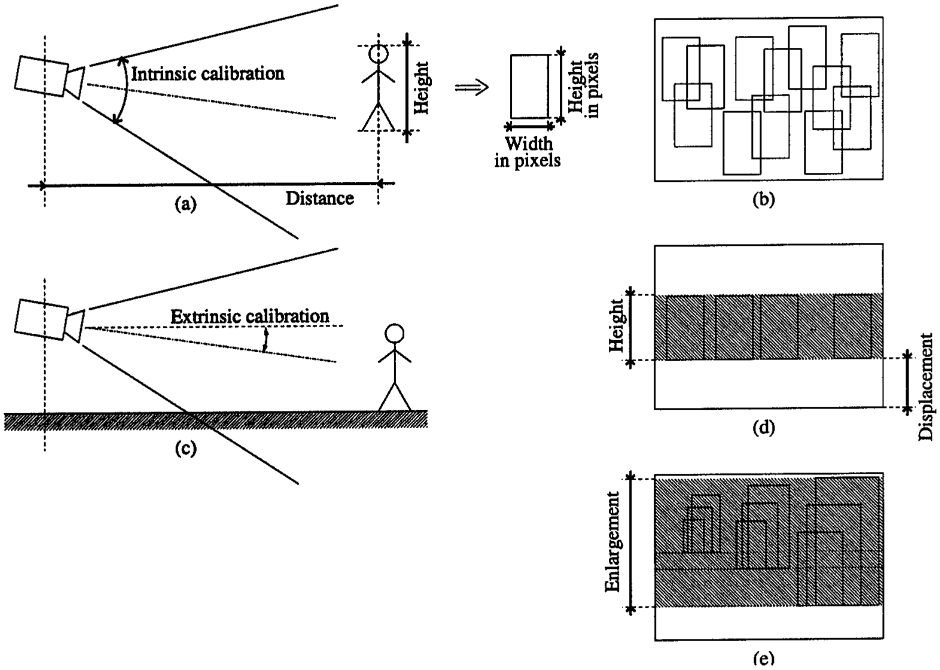


Figure 1: *(a)* Computation of the bounding box size given the intrinsic parameters and the size and distance of a pedestrian; *(b)* exhaustive search for candidates in the whole image; *(c)* the search area can be limited to a stripe given the extrinsic parameters and a flat scene assumption; *(d)* the displacement and height of the stripe depend on the pedestrian distance and height, respectively; *(e)* the search area is enlarged to explore a range of distances and heights.

3.1 Attentive vision

As a first processing step, attentive vision techniques are applied to concentrate the analysis on specific regions of interest only. In fact, the aim of the low-level part of the processing is the focusing on potential candidate areas to be further examined at a higher-level stage in the following steps.

The areas considered as candidate are rectangular bounding boxes which:

- have a size in pixels deriving from the knowledge of the intrinsic parameters of the vision system (angular aperture and resolution); in other words, once defined the size and distance of a pedestrian in the 3D world (e. g. 1.8 $m \times 0.6$ m at 20 m), simple perspective considerations give the size in pixel of its projection in the image (see figure 1.a);
- enclose a portion of the image which exhibits the low-level features that characterize the presence of a pedestrian, i. e. a strong vertical symmetry and a high density of vertical edges.

These bounding boxes will be then checked against a human shape model, taking into account the contour of the object they enclose, in order to be validated.

The search for candidates would require an exhaustive search in the whole image (see figure 1.b). However, the knowledge of the system's extrinsic parameters, together with a flat scene assumption (see figure 1.c), is exploited to limit the analysis to a stripe of the image (hereinafter referred to as *search area*). The displacement of this stripe depends on the pedestrian's distance, while its height is related to the pedestrian's height (see figure 1.d). Since by definition a pedestrian is a human shaped road participant, the flat world assumption becomes an assumption on the road slope, which is anyway a loose hypothesis in a road environment, particularly in the area immediately ahead of the vehicle. Besides the obvious advantage of avoiding false detections in wrong areas, the processing of the search area only reduces the computational time. Indeed,

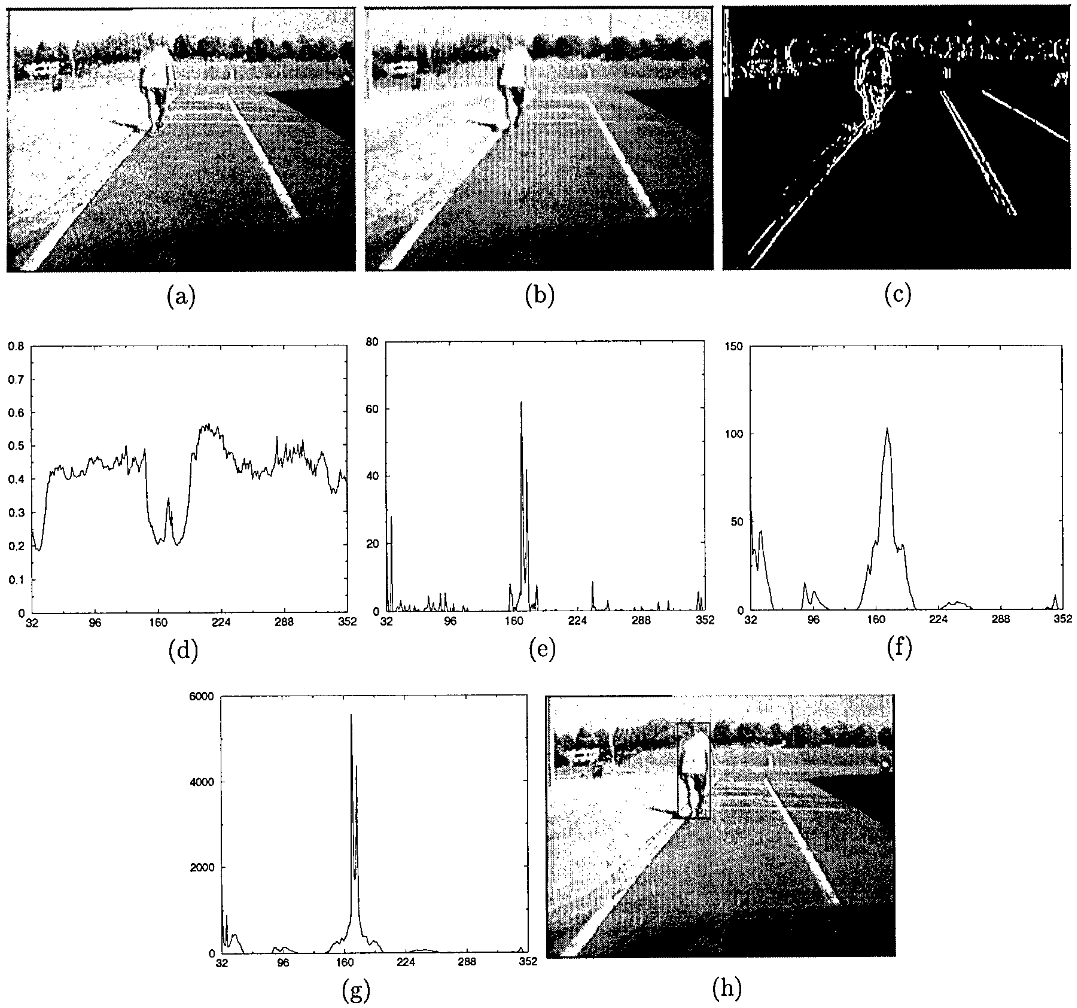


Figure 2: Intermediate results leading to the localization of bounding boxes: *(a)* original image; *(b)* clusterized image; *(c)* vertical edges; *(d)* histogram representing grey level symmetries; *(e)* histogram representing vertical edges symmetries; *(f)* histogram representing vertical edges density; *(g)* histogram representing the overall symmetry S for the best bounding box for each column; *(h)* the resulting bounding box.

the analysis cannot be limited to a fixed size and distance of the target and a given range for each parameter is in fact explored (e. g. $1.6 \div 2.0\ m \times 0.5 \div 0.7\ m$ at $10 \div 30\ m$). The introduction of these ranges generates two further degrees of freedom in the size and position of the bounding boxes. In other words, the search area is enlarged to accommodate all possible combinations of height, width, and distance (see figure 1.e).

The analysis proceeds in this way: the columns of the image are considered as possible symmetry axes for bounding boxes. For each symmetry axis different bounding boxes are evaluated scanning a specific range of distances from the camera (the distance determines the position of the bounding box base) and a reasonable range of heights and widths for a pedestrian (the corresponding bounding box size can be computed through the calibration).

However, not all the possible symmetry axes are considered: since edges are chosen as discriminant in most of the following analysis, a pre-attentive filter is applied, aimed at the selection of the areas with a high density of edges. In particular, for each axis the count of edge pixel is computed in a portion of the search area centered on the axis itself and as wide as the maximum bounding box width. Axes centered on regions

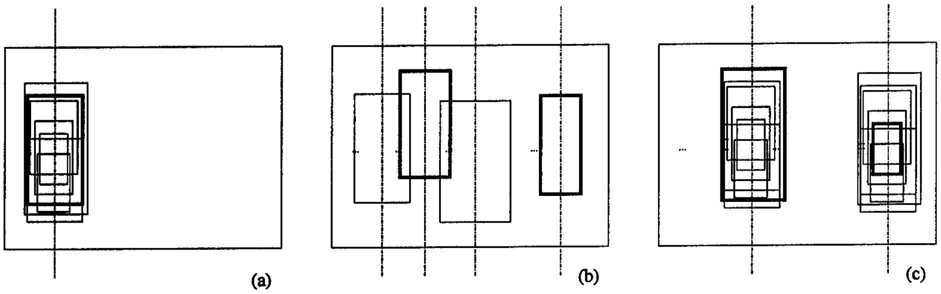


Figure 3: *(a)* Selection of the best bounding box for each symmetry axis; *(b)* selection of the best symmetry axes; *(c)* selection of the best candidates for each selected axis by choosing the bounding box which maximizes the symmetry and density of vertical edges.

which contain a number of edges lower than the average value are then dropped.

For each of the remaining axes the best candidate area is selected among the bounding boxes which share that symmetry axis, while having different position (base) and size (height and width). Vertical symmetry has been chosen as a main distinctive feature for pedestrians. Symmetry edge maps, e. g. the Generalized Symmetry Transform (GST), have already been proposed as methods to locate interest points in the image prior to any segmentation or extraction of context-dependent information. Unfortunately, these methods are generally computationally expensive. Alternatively, two different symmetry measures are performed: one on the gray-level values (G) and one on the gradient values, considering only edges with a vertical direction (E). The selection of the best bounding box is based on maximizing a linear combination of the two symmetry measures, masked by the density of edges in the box (D), as shown in the following equation: $S = (k1 \times G + k2 \times E) \times D$. The weights $k1$ and $k2$ were experimentally determined analyzing a large number of images. Figure 2 shows the original input image, the result of a clustering operation used to improve the detection of edges, a binary image containing the vertical edges, and a number of histograms representing the maximum (i) symmetry of gray-levels, (ii) symmetry of vertical edges, and (iii) density of vertical edges among the bounding boxes examined for each axis. The histogram in figure 2.g represents the linear combination of all the above. The histograms are actually computed only for the regions of the search area with a high density of edges, however in figure 2 they are completely displayed for a better understanding. It is evident that, using the density of vertical edges as a mask, interesting areas present high values for both the symmetry of gray-levels and symmetry of vertical edges. The resulting histogram is therefore thresholded and its overthreshold peaks are selected as representing candidate bounding boxes.

An adjustment of the bounding boxes' size is yet needed. In fact, when comparing the gray-level symmetry of different bounding boxes centered on the same axis, larger boxes tend to overcome smaller ones since pedestrians are generally surrounded by homogeneous areas such as concrete underneath or the sky above (this is true for other objects, too). Therefore, the bounding box which presents the maximum symmetry tends to be larger than the object it contains because it includes uniform regions. For this reason, given a peak of the overall histogram representing a selected symmetry axis, the exact height and width of the best bounding box are actually taken as those possessed by the box which maximizes a new function among the ones having the same axis. This function is computed as the product of the symmetry of vertical edges (E) and density of vertical edges (D) only. The attentive phase is sketched in figure 4.

The result of this low-level processing is a list of candidate bounding boxes which will be fed to the following stage, whose task is their validation as pedestrians, based on morphological characteristics. Prior to this higher-level validation, the list of boxes possibly surrounding human shapes is examined through a set of more detailed and specialized criteria in order to achieve a *rating* for each box. These criteria are aimed at evaluating characteristics depending on its content and this should lead to a penalty if the box is considered **not to contain** a human shape. Each box is therefore assigned a *confidence vote* which expresses its attitude towards representing a human being. This rating is then compared to a threshold and, depending on the comparison, the box can be discarded as a false positive. The criteria identified are mainly oriented

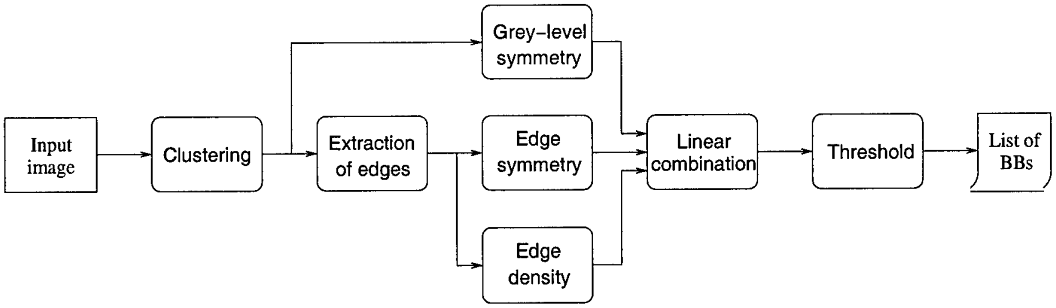


Figure 4: Sketch of the attentive phase.

to penalize vertical artifacts such as poles and trees, and boxes located in the area between two standing persons due to the high symmetry of the hole between them. They are based on:

- the distribution of edges in the bounding box, more specifically the column-wise histogram of edges is analyzed;
- the collinearity of edges, in particular trying to detect vertical lines of edges;
- the presence of wide uniform areas in the box (in this case the column-wise histogram of gray-levels is analyzed);
- the overlapping between boxes with different ratings.

3.2 Shape detection using autonomous agents

This section describes the shape detection technique.

Different edges are selected and connected, where possible, in order to form a contour. Thanks to the way the contour is built, it will represent the shape of the pedestrian body. Matching techniques may be used in regions of the bounding box that lie in the correct position in relation to the formulated hypothesis of the pedestrian.

Essentially, the process consists in adapting a deformable coarse model to the bounding boxes. Thanks to its roughness the model is sufficiently general and can be adapted to a variety of postures. Anyway, it is limited to standing pedestrians. The model adjustment is done through an evolutionary approach with a number of independent agents acting as edge trackers. The agents explore a feature map displaying the edges contained in a given bounding box and stochastically build hypotheses of a feasible contour of a human body. The idea is taken from the Ant Colony Optimization (ACO) metaheuristic devised to solve hard combinatorial optimization problems, originally inspired by the communication behavior of real ants. The system proposed here is a transposition to image analysis of one of the first ACO algorithms, the AS-*cycle*.

In nature, when ants look for food, they communicate the path and the outcome of their exploration to other ants by marking their path with a pheromone trail, its intensity depending on the distance of the food from the nest, and on its quality and quantity. Other ants are attracted by strong pheromone trails, thus the path to an abundant food source close to the nest is marked again and again until it becomes more frequented and even more attractive.

This concept can be applied to the analysis of an image by creating a colony of artificial ants that looks for an optimal combination of edge pixels that maximizes the coherency of their position according to a given model. Each ant in turn traces a solution in a solution space made up of all the possible paths connecting two pixels in a matrix. The decisional basis for each step of an ant is provided by two factors: one is a local heuristic η_i that quantifies the attractiveness of pixel i for its intrinsic characteristics; the second is the information on that pixel made available by previous attempts of other ants, in the form of a quantity of pheromone τ_i.

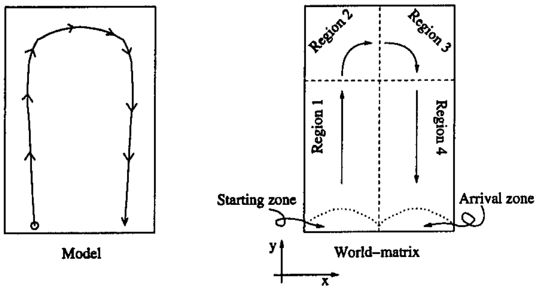


Figure 5: Artificial ants move through the world-matrix starting from the left half of the lower border, and moving through regions 1, 2, 3 and 4 until they reach the arrival line.

Artificial ants explore a world which is a matrix of pixels derived by the resampling of the edge map of the bounding box under analysis. In our experiments, the normalized world-matrix is sized 20 × 45 pixels. Each pixel i is initialized with a binary value: 1 if it contains an edge, 0 if not. This value represents its intrinsic attractiveness η_i and is the basis for the heuristic research. All the pheromone τ_i is initialized at 0. The world-matrix is visited by M ants in parallel, and the process is repeated for C cycles. At the end of each cycle, new pheromone is deposed on the trails pursued by the ants, and some of that accumulated evaporates. In this way, solutions built several cycles before, progressively loose their importance. On the other hand, pheromone on pixels that compose the path of frequently selected solutions grows. and eventually this information surpasses that given by the heuristic.

A crucial point to be understood is that artificial ants do not need to reach an optimal solution in the edge connection problem. Often, no real optimal solution exists even to a human inspector. The colony needs only to find a sufficiently valuable path that permits to continue the recognition, free of noisy edge pixels.

The ant system develops from a very elastic and deformable coarse model of a human body. The model is encoded in the progression rules that guide the ants through the solution space. The rules effectively restrict the whole space of possible solutions to a subspace that includes the searched shape. The system will then provide attempts to find a feasible path in this subspace, and each attempt will be evaluated by a confidence function as it is detailed in the following.

All ants start from the left half of the lower side of the world-matrix. The world is divided into four regions, as detailed in figure 5. In each region ants proceed of one step forward in the direction of one of the axis (y in regions 1 and 4, and x in regions 2 and 3), and can choose among a set of s pixels lying on the line or column in front of them. Additionally, in regions 2 and 3 ants have the option of moving vertically, thus they can follow very steep edges as well as very flat ones.

The starting point of each ant is chosen randomly among the edge pixels lying in the starting region, i.e. the left half of the lower border. If no edge pixel is present, the starting point is set on a random point belonging to the region. The choice of starting from edge pixels does not pose a hard restriction to the exploration of the solution space: the bounding box usually comprises edges in the lowest line owing to the mechanism that determines its dimensions based on the edge density. Most of the times, the edges appearing on the lower line of a well-centered bounding box correspond to the feet of the pedestrian. Each ant stops its journey when it reaches the right half of the lower border of the world-matrix.

An ant is an independent pixel-sized agent; it has a local exploratory capability, limited to the set of pixels belonging to the scanning region N as described above, and of those lying on the following line as well. Figure 6 illustrates the situation for the scanning sets for each region of the world-matrix. Each pixel under consideration is associated to a quality measure that takes into account terms pertaining to both the feature map of the edges, and the pheromone deposed by previous ants. The quality of pixel j is expressed as $q_j = \alpha\tau_j + (1-\alpha)\eta_j$ where η_j represents the binary heuristic information, τ_j is the quantity of pheromone accumulated at position j, and α is a parameter which determines the relative influence of the pheromone trail and the heuristic information.

Each ant always moves into one of the pixels of the nearest scanning line (line A, namely the shaded

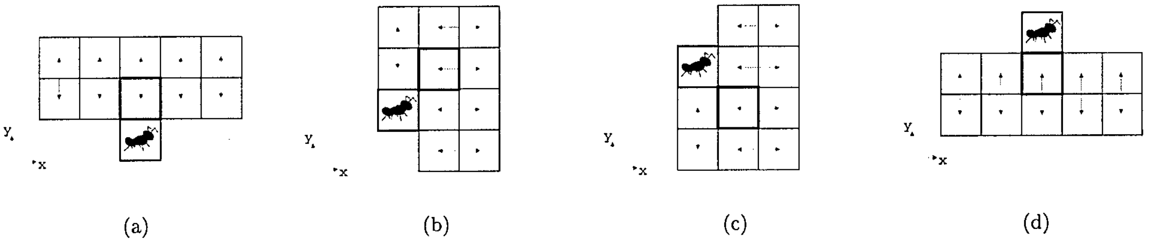


Figure 6: Artificial ants move to one pixel of the shaded set (named *line A*) by calculating the quality of each pixel of line A and of the white region (named *line B*). The figures illustrate the set of pixels evaluated by ants when they cross region 1 (a), region 2 (b), region 3 (c), and region 4 (d).

sets in figure 6), but the probability of transition combines the quality of each pixel in line A with that of a corresponding pixel in line B as indicated by the arrows in figure 6. Defining with l a pixel in line B corresponding to a pixel j in line A, the probability that ant k moves from position i to position j belonging to its feasible neighborhood N_i^k at step t is

$$p_{ij} = \frac{\frac{1}{d_j} \times [(1-\nu) \times q_j + \nu \times q_l]}{\sum\limits_{(j,l) \in N_j^k} \frac{1}{d_j} \times [(1-\nu) \times q_j + \nu \times q_l]} \tag{1}$$

where ν is a parameter in a range $[0, 1]$ indicating the ants field of view; for $\nu = 1$ the ant sees only line A pixels, for $\nu = 0$ the ant sees only line B pixels, while for intermediate values the ant focus of attention varies in between line A e B. d_j is the displacement of pixel j with respect to the central pixel of line A. The $1/d_j$ penalty favors straight trails in comparison with frequent small alternative leaps to the left and the right.

The system provides two different kinds of agents: purely stochastic ants and semi-deterministic ants. Both kinds choose their move with a uniformly distributed random rule, but the range of choice is different: purely stochastic ants have all the feasible neighborhood N illustrated in figure 6 at their disposal, while semi-deterministic ants choose only between the two pixels that have the highest p_{ij}. Both kinds of ants perform well on synthetic images; however, stochastic ants explore more widely the solution space but converge more slowly to a final solution than the semi-deterministic ants do. On the other hand, semi-deterministic ants follow well connected edges, but sometimes fail to find the best solution subspace in very irregular real images.

Once every ant has completed its tour, pheromone trails are updated through evaporation and reinforcement according to the following equation:

$$\tau_i(c+1) = (1-\rho) \times \tau_i(c) + \rho \times \left(\sum_{k=1}^{M} \Delta\tau_i^k + \Delta\tau_i^d \right) \tag{2}$$

where ρ is the evaporation coefficient (ranging from 0 to 1), $\tau_i(c)$ is the quantity of pheromone present on pixel i at cycle c. Pheromone update $\Delta\tau$ is made up of two contributions. The first one is given by the sum of the pheromone deposed by each ant at the end of its tour. The second one is credited to the best trail according to an elitist strategy.

All ants are ranked according to the following rule: an ant obtains a high rank if it takes a long tour that passes through many edges or a low rank if it visits many pixels that are not edges. This rule is functional to the search of a good solution as it encourages ants to take the shortest path between two zones of connected pixels and does not pose any request on the total length of the trail.

The procedure described above is repeated for a number of cycles; experiments show that with 10 ants, 2 cycles are sufficient for a stable and reliable solution.

Finally, the output is the path of the ant of the highest rank in the last cycle.

4 Results and discussion

This section presents the results obtained during the contract. The main aim was to develop the low-level part of a human shape localization system, but also a medium-level filter has been developed based on an innovative evolutive system. This last filter was first tested on synthetic images demonstrating a good performance. When applied to real images it showed very good performance only in cases of:

- a bounding box correctly framing a well contrasted human shape,
- a bounding box framing a manifestly wrong object, such as symmetrical road infrastructure.

For this reason this filter may be used as a detector for manifestly strong false positives. The following section will discuss the qualitative and quantitative results of the low-level phase.

4.1 Results of low-level attentive vision

The algorithm has been tested on a large number of images acquired in different situations ranging from simple uncluttered scenes to complex scenarios. As an example, figure 7 shows the results of the selection of candidate bounding boxes in three different situations. In figure 7.a a correct detection of two pedestrians is displayed. Figure 7.b presents a complex scenario in which only the central pedestrian is detected; two other pedestrians are missed because (i) the first is confused with the background, and (ii) the second is only partially visible; moreover, the tree on the right side, which could have been erroneously detected due to its high vertical symmetry, has been discarded as a false positive. In figure 7.c the pedestrian on the left side has been localized even if it is very close to the car, but two other false positives, which couldn't be discarded by the filters, are detected as well.

Some general considerations can be drawn on the behavior of this candidate selection procedure. In situations in which pedestrians are sufficiently contrasted with respect to the background and completely visible (i. e. not occluded by other pedestrians or objects) the localization of pedestrians based on symmetry and edge density proves to be robust. Thanks to the use of vertical edges, the width of the bounding boxes enclosing pedestrians is generally determined with a good precision. This allows to determine the angle under which the human shape is seen with the same good precision. On the other hand, a lower accuracy is obtained for the localization of the top and bottom of the bounding box. Since the distance to the pedestrian can be computed from the bottom of the bounding box (under the assumption that the terrain framed in the image is flat), the result of distance computation would be poor. Stereo techniques will allow to increase the distance precision without the need of the flat terrain assumption.

Symmetrical objects other than pedestrians may happen to be detected as well. In order to get rid of such false positives a number of filters have been devised which rely on the analysis of the distribution of edges within the bounding box. These filters show promising results regarding the elimination of both

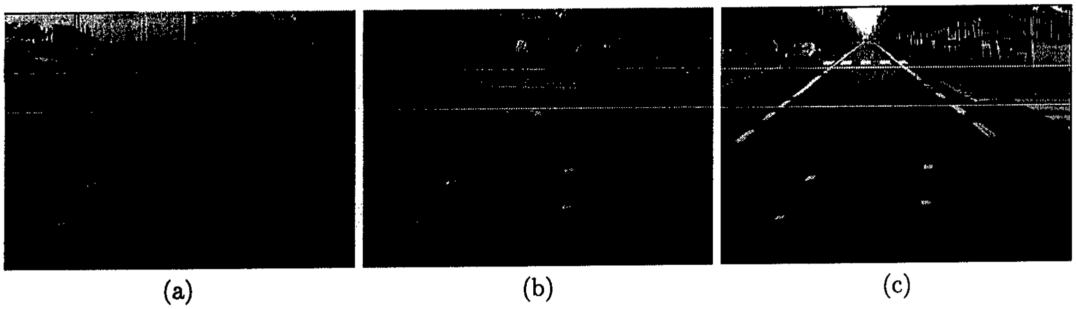
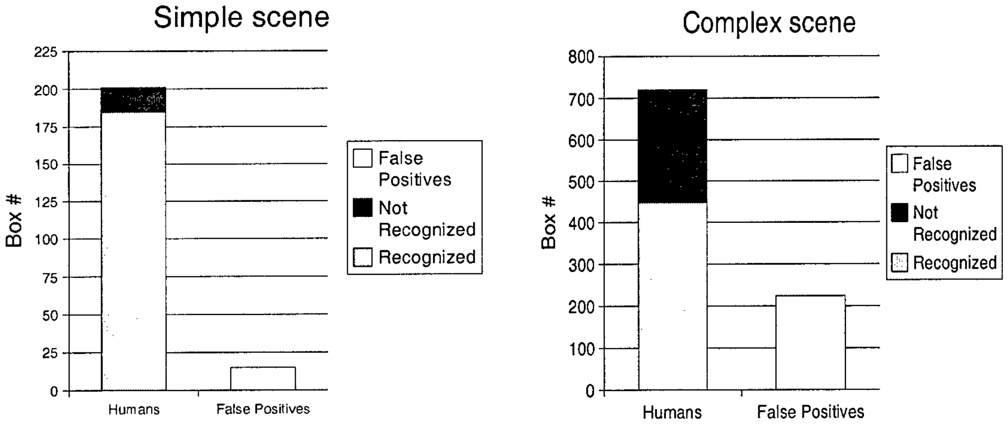

(a) (b) (c)

Figure 7: Result of low-level processing in different situations (the two white lines delimit the distance search range): *(a)* a correct detection of two pedestrians *(b)* a complex scenario in which only the central pedestrian is detected; *(c)* a crossing pedestrian has been localized, but other symmetrical areas are highlighted as well.



(a) (b)

Figure 8: Count of successes and failures of the algorithm: *(a)* excellent identification results in a simple sequence (190 images), *(b)* lower human shape identification rate in a complex sequence (801 images).

artifacts (such as poles, road signs, buildings, and other road infrastructures) and symmetrical areas given by a uniform portion of the background between two foreground objects with similar lateral borders.

A quantitative analysis of the system behavior has also been performed. The performance of the system was evaluated in the form of the percentage of correctly detected human shapes and the number of false positive detections.

The behavior of the system in situations with a different degree of complexity has been analyzed. Figure 8 summarizes the count of successes and failures of the algorithm for two sample image sequences: the former was acquired in a simple uncluttered and flat scene where only two people move and a small number of other objects are present; the latter was acquired in a complex scenario with trees and bushes where other road participants, such as vehicles and bicycles, make the localization of pedestrians more challenging. Moreover, in the first case the vehicle was still, while in the second one it was moving forward and backward. It can be observed that when in the scene there are no complex features producing a variegated environment, the correct recognition percentage rises up to 92%, while last step filters keep false positives to the low number of 15 occurrences in 190 images (see figure 8.a). Conversely, when the presence of different elements makes the scene more complex, false positives have a stronger presence (224 occurrences in 801 images), and the percentage of correct detections gets reduced to 62% since the shape of the pedestrian often merges with the background or other objects (see figure 8.b).

A specific analysis has also been performed on the behavior of the filters designed to remove false positives (see figure 9). Their introduction led to a 30% reduction of false positives (from 748 to 224 in the complex sequence). On the other hand, the percentage of correctly recognized humans is reduced from 81% to 62%. Anyway, the possibility of recovering missed pedestrians may be offered by *tracking*. In fact, though the introduction of the filters led to an increase in the maximum of consecutive false negatives associated to the same person, the mean value of the persistence of these misdetections was reduced, as can be observed in table 1.

Therefore, as a final consideration, the introduction of stereo techniques associated to a tracking phase would allow to improve the overall system.

A List of publications produced

The following publications were produced with the results achieved during the contract.

- Massimo Bertozzi, Alberto Broggi, Massimo Cellario, Alessandra Fascioli, Paolo Lombardi, and Marco Porta, Artificial Vision in Road Vehicles, Proceedings of the IEEE - Special issue on "Technology and

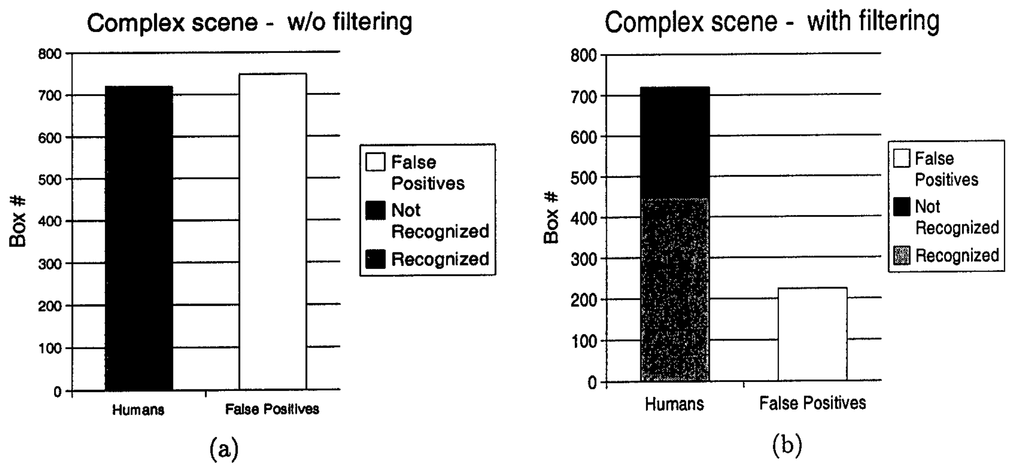


Figure 9: Behavior of the algorithm on the same complex sequence depicted in the previous figure: *(a)* without filters, *(b)* after the application of filters.

Tools: Visual Perception", July 2002.

- Massimo Bertozzi, Alberto Broggi, Alessandra Fascioli, and Paolo Lombardi, Vision-based Pedestrian Detection: will Ants Help?, In Procs. IEEE Intelligent Vehicles Symposium 2002, Paris, France, June 2002.
- Massimo Bertozzi, Alberto Broggi, Alessandra Fascioli, Paolo Lombardi, and Amos Tibaldi, Vision-based Pedestrian Detection in Road Environments, 2nd Annual Intelligent Vehicle System Symposium, Traverse City, MI, USA, June 3 – 5, 2002.

B List of scientific personnel

The following people contributed to the research during the contract:

- Principal Investigator: Prof. Alberto Broggi (professor)
- Research Associate: Prof. Gianni Conte (full professor)
- Research Associate: Dr. Massimo Bertozzi (researcher)
- Research Assistant: Dr. Alessandra Fascioli (PhD)
- Research Assistant: Dr. Ugo Vallone (PhD Candidate)
- Research Assistant: Dr. Paolo Lombardi (PhD Candidate)
- Research Assistant: Dr. Amos Tibaldi (PhD Candidate)

	N. of images	Maximum	Mean
without filters	801	12	2.96
with filters	801	22	2.64

Table 1: Consecutive false negatives for the same person.

Artificial Vision in Road Vehicles

M. Bertozzi*, A. Broggi*, M. Cellario°, A. Fascioli*, P. Lombardi°, M. Porta°

*Dipartimento di Ingegneria dell'Informazione
Università di Parma
Parco Area delle Scienze, 181A
I-43100 PARMA, Italy

°Dipartimento di Informatica e Sistemistica
Università di Pavia
Via Ferrata, 1
I-27100 PAVIA, Italy

Contact author:
Massimo Cellario
E-mail: cellario@vision.unipv.it

Abstract

The last few decades witnessed the birth and growth of a new sensibility to transportation efficiency. In particular, the need for efficient and improved people and goods mobility pushed researchers to address the problem of intelligent transportation systems. This paper surveys the most advanced approaches to the (partial) customization of road following task, using on-board systems based on artificial vision. The functionalities of Lane Detection, Obstacle Detection and Pedestrian Detection are described and classified, and their possible application on future road vehicles is discussed.

Key words

Intelligent Transportation Systems (ITS); Automatic Vehicles Guidance; Intelligent Vehicles; Machine Vision

1. Introduction

Problems concerning traffic mobility, safety, and energy consumption have become more serious in most developed countries in recent years. The endeavors in solving these problems have triggered the interest towards new fields of research and application, such as Automatic Vehicle Driving, in which new techniques are investigated for the entire or partial automation of driving tasks. A recently defined comprehensive and integrated system approach, referred to as Intelligent Transportation Systems (ITS), links the vehicle, the infrastructure and the driver to make it possible to achieve more mobile and safer traffic conditions by using state-of-the-art electronic communication and computer-controlled technology.

Over time, ITS research community expects that intelligent vehicles will advance in three primary ways: in the capabilities of in-vehicle systems, in the sophistication of the driver-vehicle interface, and in the ability of vehicles to communicate with each other and a smart infrastructure [1].

Smart vehicles will be able to give route directions, sense objects, warn drivers of impending collisions, automatically signal for help in emergencies, keep drivers alert, and may ultimately be able to take over driving.

In fact, ITS technologies may provide vehicles with different types and levels of "intelligence" to complement the driver. Information systems expand the driver's knowledge of routes and locations. Warning systems, such as collision-avoidance technologies, enhance the driver's ability to sense the surrounding environment. Driver assistance and automation technologies simulate the driver's sensor-motor system to operate a vehicle temporarily during emergencies or for prolonged periods.

The timing of "human-centered" intelligent vehicles' arrival on the market, however, will depend on the resolution of technical and cost constraints for some advanced concepts, such as collision-avoidance and automated systems, manufacturers' interest, production lead-times, and consumer demand.

Human-centered intelligent vehicles hold a major potential for industry. Since 1980, major car manufacturers and other firms began developing computer-based in-vehicle navigation systems. Today, most developed/developing systems around the world have included more complex functions to help people drive their vehicles safely and efficiently.

New information and control technologies that make vehicles smarter are now arriving on the market either as optional equipment or as specialty after-market components. These technologies are being developed and marketed to increase driver safety, performance, and convenience. However, these disparate individual components have yet to be integrated to create a coherent intelligent vehicle that complements the human driver, fully considering his requirements, capabilities and limitations.

A fully intelligent vehicle must work cooperatively with the driver [1]: an intelligent system senses its environment and acts to reach its objectives: its interaction-communication channels have a big influence on the type of intelligence it can display [2].

New uncoordinated technologies could deliver excessive, competing, or contradictory messages and demands that might distract, confuse and overwhelm the driver, overloading his limited cognitive resources and eventually leading to a decrease in his own performance and safety.

Clearly there is a need in the research community for quantitative and objective performance metrics to define and structure this problem domain, for describing and evaluating products in future competitive markets when research is reduced to technology and commercialized.

In the last two decades, governmental institutions have activated initial explorative phases by means of various projects worldwide, involving a large number of research units who worked in a cooperative way producing several prototypes and solutions, based on rather different approaches.

In Europe the PROMETHEUS project (PROgraM for a European Traffic with highest Efficiency and Unprecedented Safety) started this explorative stage in 1986. The project involved more than thirteen vehicle manufacturers and several research units from governments and universities of nineteen European countries. Within this framework, a number of different ITS approaches were conceived, implemented, and demonstrated.

In the United States a great deal of initiatives were launched to address the mobility problem, involving universities, research centers, and automobile companies. After this pilot phase, in 1995 the US government established the National Automated Highway System Consortium (NAHSC) [3], and launched the Intelligent Vehicle Initiative (IVI) right after in 1997.

In Japan, where the mobility problem is even more intense and evident, some vehicle prototypes were also developed within the framework of different projects. Similarly to the US case, in 1996 the Advanced Cruise-Assist Highway System Research Association (AHSRA) was established amongst a large number of automobile industries and research centers [4], which developed different approaches to the problem of Automatic Vehicle Guidance.

As a whole, the main results of this first stage provided a deep analysis of the problem and the development of a feasibility study to understand the requirements and possible effects of ITS technology applications.

The ITS field is now entering its second phase characterized by a maturity in approaches and by new technological possibilities which allow the development of the first experimental products. A number of prototypes of intelligent vehicles have been designed, implemented, and tested on the road. The design of these prototypes has been preceded by the analysis of solutions deriving from similar and close fields of research, and has produced a great flourishing of new ideas, innovative approaches, and novel ad hoc solutions. Robotics, artificial intelligence, computer science, computer architectures, telecommunications, control and automation, signal processing are just some of the principal research areas from which the main ideas and solutions were first derived. Initially, underlying technological devices - such as head-up displays, infrared cameras, radars, sonars - derived from expensive military applications, but, thanks to the increased interest in these applications and to the progress in industrial production, today's technology offers sensors, processing systems, and output devices at very competitive prices. In order to test a wide spectrum of different approaches, these automatic vehicles prototypes are equipped with a large number of different sensors and computing engines.

Section 2 of this paper describes the motivations which underlie the development of vision-based intelligent vehicles, and illustrates their requirements and peculiarities. Section 3 surveys the most common approaches to Road Following developed worldwide; Section 4 briefly introduces a few significant architectural issues, while Section 5 outlines our perspectives in the evolution of intelligent vehicles.

2. Improving On-road Mobility

2.1 Intelligent vehicles vs. intelligent infrastructures

The enhancement of future vehicles' efficiency can be achieved acting both on infrastructures and on vehicles. Depending on the specific application, either choice possesses advantages and drawbacks. Enhancing road infrastructure may yield benefits to transportation architectures based on repetitive and prescheduled routes, such as public transportation and industrial robotics. On the other hand, extended road networks for private vehicles would require a complex and extensive organization and maintenance which may become cumbersome and extremely expensive: an ad hoc structuring of the environment can only be considered for a reduced subset of the road network, for example a fully automated highway on which only automatic vehicles -public or private- can drive.

A great deal of research has been focused on advancement in vehicle safety and automation systems, applied to different major classes of vehicles: light, commercial, transit, and specialty vehicles; particular attention has been dedicated to selected problem areas, as "prime candidates" for improving vehicle safety and efficiency: lane keeping and road departure warning, collision avoidance (intersection, merge, rear-end, vehicles, obstacles, pedestrians), vision enhancement, vehicle stability, driver condition monitoring and safety-impacting in-vehicle technology integration.

In this paper only in-vehicle control and automation applications are considered, while road infrastructure, inter-vehicle communication, satellite communication, information systems, and driver-vehicle interface issues are not covered.

Any on-board system for ITS applications needs to meet some important requirements:

- the final system, installed on a commercial vehicle, must be robust enough to adapt to different conditions and changes of environment, road, traffic, illumination, and weather. Moreover, the hardware system needs to be resistant to mechanical and thermal stress.
- On-board systems for ITS applications are safety critical and require a high degree of reliability: the project has to be thorough and rigorous during all its phases, from requirements specification to design and implementation. An extensive phase of testing and validation is therefore of paramount importance.
- For marketing reasons, the design of an ITS system is driven by strict cost criteria (it should cost no more than 10% of the vehicle price) thus requiring a specific engineering phase. Operative costs (such as power consumption) need to be kept low as well, since vehicle performance should not be affected by the use of ITS apparata.
- System's hardware and sensors have to be kept compact in size and should not disturb car styling.
- The design of the driver-vehicle interface (the place where the driver interacts physically and cognitively with the vehicle) is critical. When giving drivers access to ITS systems inside the vehicle, designers must not only consider safety (i.e., overloading the driver's information-processing resources), but also usability and driver acceptance [5]: interfaces will need to be intelligent and user-friendly, effective and transparent to use; in particular, a full understanding of the subtle trade-offs of multimodal interface integration will require significant research [2].

2.2 Active vs. passive sensors

Among the sensors widely used in indoor robotics, tactile sensors and acoustic sensors are of no use in automotive applications because of vehicles' speed and reduced detection range.

Laser-based sensors and millimeter-wave radars detect the distance of objects by measuring the travel time of a signal emitted by the sensors themselves and reflected by the object, and are therefore classified as *active sensors*. Their main common drawbacks consist in low spatial resolution and slow scanning speed. However, millimeter-wave radars are more robust to rain and fog than laser-based radars, though more expensive.

Vision-based sensors are defined as *passive sensors* and have an intrinsic advantage over laser and radar sensors: the possibility of acquiring data in a non-invasive way, thus not altering the environment (image scansion is performed fast enough for ITS applications). Moreover, they can be used for some specific applications for which visual information plays a basic role (such as lane markings localization, traffic signs recognition, obstacle identification) without requiring any modifications to road infrastructures. Unfortunately vision sensors are less robust than millimeter-wave radars in foggy, night, or direct sun-shine conditions.

Active sensors possess some specific peculiarities which result in advantages over vision-based sensors, in this specific application: they can measure some quantities, such as movement, in a more direct way than vision and require less performing computing resources, as they acquire a considerably lower amount of data.

Nevertheless, besides the problem of environment pollution, the wide variation in reflection ratios caused by different reasons (such as obstacles shape or material) and the need for the maximum signal level to comply with some safety rules, the main problem in using active sensors is represented by interference among sensors of the same type, which could be critical for a large number of vehicles moving simultaneously in the same environment, as -for example- in the case of autonomous vehicles traveling on intelligent highways.

Hence, foreseeing a massive and widespread use of autonomous sensing agents, the use of passive sensors, such as cameras, obtains key advantages over the use of active ones.

Obviously machine vision does not extend sensing capabilities besides human possibilities in very critical conditions (e.g., in foggy weather or during the night with no specific illumination), but can, however, help the driver in case of failure, for example in the lack of concentration or drowsy conditions.

2.3 Vision-based intelligent vehicles

Some important issues must be carefully considered in the design of a vision system for automotive applications. In the first place, ITS systems require faster processing than other applications, since vehicle speed is bounded by the processing rate. The main problem that has to be faced when real-time imaging is concerned and which is intrinsic to the processing of images, is the large amount of data -and therefore computation- involved. As a result, specific computer architectures and processing techniques must be devised in order to achieve real-time performance. Nevertheless, since the success of ITS apparata is tightly related to their cost, the computing engines cannot be based on expensive processors. Therefore, either of-the-shelf components or ad hoc dedicated low-cost solutions must be considered.

Secondly, in the automotive field no assumptions can be made on key parameters, for example scene illumination or contrast, which are directly measured by the vision sensor. Hence, the subsequent processing must be robust enough to adapt to different environmental conditions (such as sun, rain, fog) and to their dynamic changes (such as transitions between sun and shadow, or the entrance or exit from a tunnel).

Furthermore, other key issues, such as the robustness to vehicle's movements and drifts in the camera's calibration, must be handled as well.

However, recent advances in both computer and sensor technologies promote the use of machine vision also in the intelligent vehicles field. The developments in computational hardware, such as a higher degree of integration and a reduction of the power supply voltage, permit to produce machines that can deliver a high computing power, with fast networking facilities, at an affordable price. Current technology allows the use of SIMD-like processing paradigms even in general-purpose processors, such as the new generation of processors that include multimedia extensions.

In addition, current cameras include new important features that permit the solution of some basic problems directly at sensor level. For example, image stabilization can be performed during acquisition, while the extension of camera dynamics allows to avoid the processing required to adapt the acquisition parameters to specific light conditions. The resolution of the sensors has been drastically enhanced, and in order to decrease the acquisition and transfer time, new technological solutions can be found in CMOS sensors, such as the possibility of dealing with pixels independently as in traditional memories. Another key advantage of CMOS-based sensors is that their integration on the processing chip seems to be straightforward.

Many different parameters must be evaluated for the design and choice of an image acquisition device. First of all, some parameters tightly coupled with the algorithms regard the choice of monocular vs. binocular (stereo) vision and the sensors' angle of view (some systems adopt a multi-camera approach, by using more than one camera with different viewing angles, e.g. fish eye or zoom). The resolution and the depth (number of bit/pixel) of the images have to be selected as well (this also includes the selection of color vs. monochrome images).

Other parameters -intrinsic to the sensor- must be considered. Although the frame rate is generally fixed for CCD-based devices (25 or 30 Hz), the dynamics of the sensor is of basic importance: conventional cameras allow an intensity contrast of 500:1 within the same image frame, while most ITS applications require a 10,000:1 dynamic range for each frame and 100,000:1 for a short image sequence. Different approaches have been studied to meet this requirement, ranging from the use of CMOS-based cameras with a logarithmically compressed dynamic [6], [7] to the interpolation and superimposition regarding values of two subsequent images taken from the same camera [8].

In conclusion, although extremely complex and highly demanding, computer vision is a powerful means for sensing the environment and has been widely employed to deal with a large number of tasks in the automotive field, , thanks to the great deal of information it can deliver (it has been estimated that humans perceive visually about 90% of the environment information required for driving).

3. The Road Following Driving Task

Among the complex and challenging tasks of future road vehicles is Road Following. It is based on: Lane Detection (which includes the localization of the road, the determination of the relative position between vehicle and road, and the analysis of the vehicle's heading direction), and Obstacle Detection (which is mainly based on localizing possible obstacles on the vehicle's path). Moreover, growing demand is also posed on the problem of Pedestrian Detection, since safety and avoidance of car crushes with pedestrians is a central concern for future systems.

In this section, a survey on the most common approaches to Lane Detection, Obstacle Detection and Pedestrian Detection is presented, focusing on vision-based systems.

3.1 Lane Detection

In most prototypes of autonomous vehicles developed worldwide, Lane Following is divided into the following two steps: initially the relative position of the vehicle with respect to the lane is computed and then actuators are driven to keep the vehicle in the correct position.

Conversely, some early systems were not based on the preliminary detection of the road's position, but obtained the commands to be issued to the actuators (steering wheel angles) directly from visual patterns detected in the incoming images. For example, the ALVINN (Autonomous Land Vehicle In a Neural Net) system is based on a neural net approach: it is able to follow the road after a training phase with a large set of images [9].

Nevertheless, since the knowledge of the lane position can be conveniently exploited by other driving assistance functions, the localization of the lane is generally performed.

A few systems have been designed to handle completely unstructured roads: for example, the SCARF (Supervised Classification Applied to Road Following [10]) and PVR III (POSTECH Road Vehicle [11]) systems are based on the use of color cameras and exploit the assumption of a homogeneously colored road to extract the road region from the images.

More generally, however, Lane Detection has been reduced to the localization of specific features such as markings painted on the road surface. This restriction eases the detection of the road, nevertheless two basic problems must be faced.

- The presence of shadows (projected by trees, buildings, bridges, or other vehicles) produces artifacts onto the road surface, and thus alterates the road texture. Most research groups face this problem using highly sophisticated image filtering algorithms. When lane markings are not well visible (because of low contrast, shadows, bad weather conditions, etc.), the use of pattern-based techniques can be helpful. The system developed at the Toyota Central R&D Labs, for example, is based on a voting method, in which lane markings patterns are generated and provided. After ordinary edge extraction, edge points are matched to each pattern. At the end of the process, the patterns with the greater number of votes are chosen as the best approximations of the left and right lane markings [12]. Other algorithms exploit the processing of color images, this is the case of the MOSFET (Michigan Offroad Sensor Fusing Experimental Testbed) autonomous vehicle which uses a color segmentation algorithm that maximizes the contrast between lane markings and road [13].

- Other vehicles on the path partly occlude the visibility of the road and therefore of road markings, as well.

  To cope with this problem, some systems have been designed to investigate only a small portion of the road ahead of the vehicle where the absence of other vehicles can be assumed. As an example, the LAKE and SAVE autonomous vehicles rely on the processing of the image portion correspondent to the nearest 12 meters of road ahead of the vehicle, and it has been demonstrated that this approach is able to safely maneuver the vehicle on highways and even on belt ways or ramps with a bending radius down to 50 meters [14].

On the other hand, some systems solve the occlusion problem by combining lane detection with obstacle detection. For example, the RALPH (Rapidly Adapting Lateral Position Handler) system reduces the portion of the image to be processed according to the result of a radar-based obstacle detection module [15].

The algorithm developed by General Dynamics Robotic Systems and The Ohio State University exploits the road gray level histogram to detect lane markings, which are then analyzed using a decision tree. A histogram-based segmentation classifies the objects in the scene as road, lane markings candidates or obstacle (vehicle) candidates [16].

In other cases, the search area for lane markings detection is determined first. The research group of the Laboratoire Central des Ponts-et-Chaussees de Strasbourg assumes that there should always be a chromatic contrast between road and off-road areas (or obstacles), at least in one color component; the concept of chromatic saturation is used to separate the components [17].

Since Lane Detection is generally based on the localization of *specific patterns* (lane markings), it can be performed with the analysis of a single still image. In addition, some assumptions may help and/or speed-up the detection process.

- Due to both physical and continuity constraints, the processing of the whole image can be replaced by the analysis of specific regions of interest only (the so-called focus of attention), in which the features of interest are more likely to be found. This is a generally followed strategy that can be adopted using the results of previously processed frames or assuming an a-priori knowledge on the road environment.

In some approaches, in particular, windows of interest (WOIs) are determined dynamically by means of statistical methods. For example, the system developed at LASMEA selects the proper window according to the current state and previously detected WOIs [18].

A system developed by the Robert Bosch GmbH research group, on the other hand, employs a model both for the road and the vehicle's dynamic to determine the road portion where it is most likely to find lane markings [19].

- The assumption of a fixed or smoothly varying lane width allows the enhancement of the search criterion, limiting the search to almost parallel lane markings.

As an example, on the PVR III vehicle, lane markings can be detected using both neural networks and simple vision algorithms: two parallel stripes of the acquired image are selected and filtered using Gaussian masks and zero crossing to find vertical edges. The result is matched against a given model (a typical road pattern with parallel lane markings) to compute a steering angle and a fitness evaluation indicating the confidence in the result [11].

Analogously, the RALPH system is based on the processing of the image portion corresponding to the road about 20 to 70 meters ahead of the vehicle, depending on the vehicle's speed and obstacles presence. The perspective effect is removed from this portion of the image and the determination of the curvature is carried out according to a number of possible curvature models for a specific road template featuring parallel road markings [15].

- The reconstruction of road geometry can be simplified by assumptions on its shape.

The research groups of the Universität der Bundeswehr [20], Daimler-Benz [21] and Robert Bosch GmbH [22] base their road detection functionality on a specific road model: lane markings are modeled as clothoids. In a clothoid the curvature depends linearly on the curvilinear reference. This model has the advantage that the knowledge of two parameters only allows the full localization of lane markings and the computation of other parameters like the lateral offset within the lane, the lateral speed with respect to the lane, and the steering angle. Another system based on a clothoidal modelization of lane markings is the one developed at The Ohio State University, where a dynamic programming optimization method is used to chose among center-line candidates representing the actual geometry of the road [23].

Other research groups use a polynomial representation for lane markings. In the MOSFET autonomous vehicle, for instance, lane markings are modeled as parabolas [13]. A simplified Hough transform is used to accomplish the fitting procedure.

Similarly, a preliminary version of the lane detection system developed at The Ohio State University Center for Intelligent Transportation Research relies on a polynomial curve [24]. It assumes a flat road with either continuous or dashed bright lane markings. The history of previously located lane markings is used to determine the region of interest, thus reducing the portion of the image to be processed. The algorithm extracts the significant bright regions from the image plane and stores them in a vector list. Qualitative parameters such as lines convergence at infinity or lane width, known or estimated, are used to extract the candidate lane markings from the list. Finally, in order to handle also dashed lines, a low-order polynomial curve is fitted to the computed vectors.

As another example, a system developed at The Université Blaise Pascal exploits a polynomial road modelization to calculate the impact distance from the vehicle to the nearest road side. Such distance is obtained by considering the intersection between the straight line trajectory followed in case of a driver loss of control and the polynomial function describing the road side [25].

Recently, a research group from the University of Michigan has proposed to use concentric circles to represent lane boundaries. Since, at least in the USA, lanes are actually laid on concentric circles, circular shape models can in fact be better choices than polynomial approximations [26].

On the contrary, other systems adopt a more generic model for the road. The ROMA vision-based system uses a contour-based method [27]. A dynamic road model permits the processing of small portions of the acquired images therefore enabling real-time performance. Actually, only straight or small curved roads without intersections are included in this model. Images are processed using a gradient-based filter and a programmable threshold. The road model is used to follow contours formed by pixels that feature a significant gradient direction value.

The CyCab electric vehicle uses an edge linking process based on Contour Chains and Causal Neighborhood Windows (areas of interest connected to edge elements). After an initial segmentation phase, the longest chains with slope angles close to 45 and 135 degrees are searched for, as they represent the most probable candidates for left and right lanes [28].

Similarly, the system implemented at the Transportation College of Jilin University of Technology is based on a linear lane model, where road markers are reconstructed as sequences of straight lines [29].

A generic triangular road model was originally tested on the MOB-LAB experimental vehicle by the research groups of the University of Parma [30] and the Istituto Elettrotecnico Nazionale "G. Ferraris", CNR, Italy [31].

- The knowledge of the specific camera calibration together with the assumption of an a-priori knowledge on the road (i.e. a flat road without bumps) can be exploited to ease the localization of features and/or to simplify the mapping between image pixels and their correspondent world coordinates.

The majority of previously discussed systems exploit the assumption of a flat road in front of the vehicle in the determination of obstacle distance or road curvature, once the specific features of interest have been localized in the acquired image. The GOLD (Generic Obstacle and Lane Detection) system [32] implemented on the ARGO autonomous vehicle and the already mentioned RALPH system, however, exploit this assumption also in the lane determination process. In fact, in both cases the lane markings detection is performed in a different image domain, representing a bird's eye view of the road, which can be obtained thank to the flat road assumption.

Table 1 summarizes the pros and cons of the assumptions on which the most common approaches to lane detection rely.

Table 1: Pros and Cons of the most typical assumptions in lane detection

	Pros	Cons
Focus of attention	Analysis of a small image portion, fast processing, real-time performance, low-cost hardware	Does not fit large features, choice of the region of interest is critical
Fixed lane width	Enhancement of the search criterion (parallel line markings), robustness to shadows and occlusions	Does not match roads with variable lane width
Road shape	Strong model robust w.r.t. shadows and occlusions, eases road geometry reconstruction, simplifies vehicle control	Requires fitting with complex equations, high computational power needed, fails if the road does not match the model
A priori knowledge on the road surface/slope	Simplifies mapping between image pixels and world coordinates (e.g., determination of obstacle distance, road curvature)	Hypotheses (e.g., flat road) are not always met in real cases, approximations of recalibration needed

Other methods, based on statistical approaches, have been experimented to cope with unfriendly lighting and weather conditions. In the already quoted system implemented at LASMEA, for instance, the search for road markers is carried out as an iterative process, where continuous updates of the lane model and the size of areas of interest allow the lane detection task to be relatively noise insensitive [33], [18].

When the processing is aimed not only at a mere lane detection but also at lane markings tracking, the temporal correlation between consecutive frames can be used either to ease the feature determination or to validate the result of the processing. The lane detection module implemented and tested on the ARGO vehicle falls in the first case as it restricts the image portion to be analyzed to the nearest neighborhood of the markings previously detected [34]. In a different manner, the lane detection module developed by the research group at the Istituto Elettrotecnico Nazionale "G. Ferraris", uses the result of previous computations to validate the current one, once lane markings are found by means of a triangular model [31].

3.2 Obstacle Detection

The criteria used for the detection of obstacles depend on the definition of what an obstacle is (see Fig. 1).

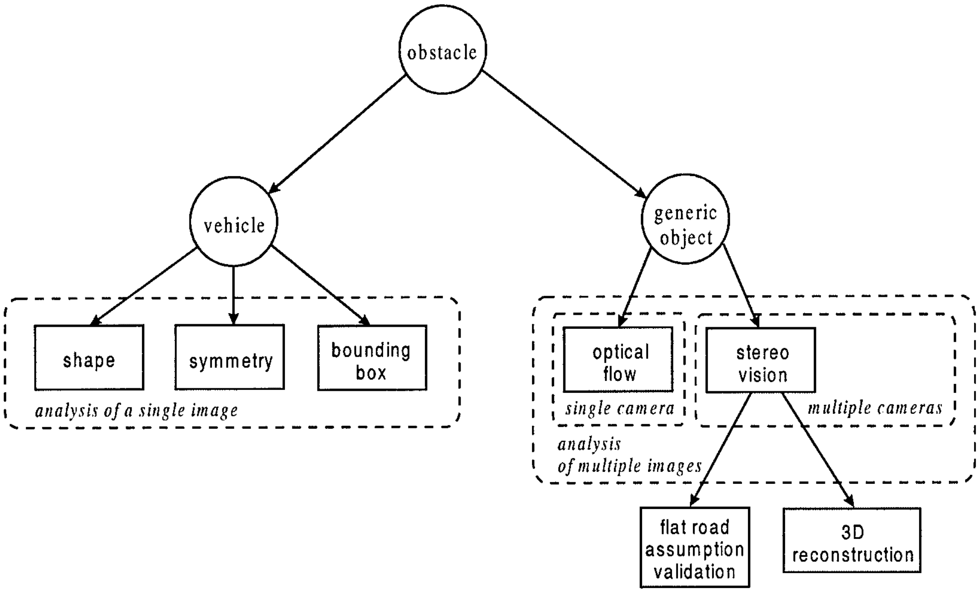


Fig. 1: Depending on the definition of obstacle, different techniques are used

In some systems determining obstacles is limited to the localization of vehicles, which is then based on a search for specific patterns, possibly supported by other features, such as shape, symmetry, or the use of a bounding box.

Conversely, the Obstacle Detection algorithm developed at the Universität der Bundeswehr is based both on an edge detection process and on obstacle modelization; the system is able to detect and track up to twelve objects around the vehicle. The continuously updated obstacle variables are: distance, direction, relative speed, relative acceleration, lateral position, lateral speed, and size [20].

Analogously, in the vehicle detecting and tracking system developed by the R&D Group at NEC Corporation, the search for possible cars and trucks is composed of two stages. After an edge-based potential vehicle identification procedure, a vehicle validation process is carried out. By exploiting characteristics such as symmetry, shadow underneath the vehicle and differences in gray level average intensities, false detections can be usually removed, also in case of bad weather conditions [35].

When Obstacle Detection is limited to the localization of specific patterns, as in the previous examples, processing can be based on the analysis of a single still image, in which relevant features are searched for. For example, in another system developed at the Istituto "G. Ferraris" a strategy is proposed which formulates obstacle hypotheses by region segmentation algorithms. The hypotheses are then validated by matching edge segmentations of these regions with a dynamic model of the vehicle, which takes into account the various rear parts of a typical car (rear window, bumper, license plate, etc.) [36].

However, unfortunately, the pattern-based approach is not successful when an obstacle does not match the model.

A more general definition of obstacle, which obviously leads to more complex algorithmic solutions, identifies as an obstacle any object that obstructs the vehicle's driving path or, in other words, anything rising out significantly from the road surface. In this case, Obstacle Detection is reduced to identifying the free-space (the area in which the vehicle can safely move) instead of recognizing specific patterns.

Due to the general applicability of this definition, the problem is dealt with using more complex techniques; the most common ones are based on the processing of two or more images, such as:

- the analysis of optical flow field and
- the processing of non-monocular images.

The optical flow-based technique requires the analysis of a sequence of two or more images: a 2D vector is computed in the image domain, encoding the horizontal and vertical components of the velocity of each pixel. The result can be used to compute ego-motion, which in some systems is directly extracted from odometry; obstacles can be detected by analyzing the difference between the expected and real velocity fields.

As an example, the ROMA system integrates an obstacle detection module that is based on the use of an optical flow technique in conjunction with data coming from an odometer [37].

Similarly, the ASSET-2 (A Scene Segmenter Establishing Tracking v2) is a complete real-time vision system for segmentation and tracking of independently moving objects. Its main feature is that it does not require any camera calibration. It tracks down objects and is capable of correctly handling occlusions amongst obstacles, and automatically tracks down each new object that enters the scene. ASSET-2 initially builds a sparse image flow field and then segmentates it into clusters that feature homogeneous flow variation. Temporal correlation is used to filter the result, therefore improving accuracy [38].

As another example, the system developed at the Kumamoto University uses a technique based on Focus of Expansion (FOE) to carry out camera 3D motion analysis. By removing background changes, moving objects can be detected and tracked in the scene. The tracking method is based on a continuous estimation of moving objects' "vitality" and "reliability" values and can deal with up to more than thirty objects in real time. Vitality of a tracked object increases when there is a sequence of template matching successes, while decreases (eventually reaching zero) after a sequence of bad matchings, indicating that the object cannot be identified further. Reliability, instead, is the quality of a template matching [39].

Still exploiting monocular vision, other research groups base their techniques on simpler principles. For example, at LASMEA a system has been developed which regulates the speed so as to respect safety distances from the preceding vehicles. In order to greatly simplify the detection task, the approach employs vehicles bearing visual marks (the rear left and right lamps and a roof lamp). From the known configuration of such visual elements, the targets can be easily located in 3D [40].

On the other hand, the processing of non-monocular image sets requires identifying correspondences between pixels in the different images: two images, in the case of stereo vision, and three images, in the case of trinocular vision. The advantage of analyzing stereo images instead of a monocular sequence lies in the possibility of directly detecting the presence of obstacles, which, in the case of an optical flow-based approach, is indirectly derived from the analysis of the velocity field. Moreover, in a limit condition where both vehicle and obstacles have small or null speeds, the optical flow-based approach fails while the other can still detect obstacles.

The UTA project (Urban Traffic Assistant) of the Daimler-Benz research group for example, aims at an intelligent stop and go for inner-city traffic using stereo vision and obtaining 3D information in real time. In addition, the UTA demonstrator is able to recognize traffic signs, traffic lights and walking pedestrians as well as the lane, zebra crossings, and stop lines [21].

Also the Massachusetts Institute of Technology group developed a cost-effective stereo vision system. The system is used for three-dimensional lane detection and traffic monitoring, as well as, for other on-vehicle applications. The system is able to separate partially overlapped vehicles and distinguish them from shadows.

As another example, the system developed by the Research & Development Center at Toshiba Corporation is based on a domain-specific stereo method for 2D navigation without depth search and metric camera calibration. Under the assumption that the vehicle is moving on a flat plane, it uses a "pseudo-projective camera model", which provides a good approximation to the general camera model in road scenes [41].

In stereovision, the correct identification of correspondences in the two images represents an important problem. In fact, the size of the areas where the search for corresponding pixels is performed can deeply influence the quality of the results obtained. If the window size is too small, the right match may be missed. On the other hand, if the window size is too large, too many possibilities may exist. To overcome this problem, some systems try to identify the image zones in which it is more probable to find homologous points. The stereo matching algorithm developed at the Tohoku University, for example, is based on SAD (Sum of Absolute Differences) computation and uses variable window sizes for each pixel in the image. Window dimensions for a specific pixel are determined by searching for minimums in the corresponding SAD graph [42].

Other systems face the stereo correspondence problem in completely different manners. For instance, the algorithm studied at the Universitè des Sciences et Technologies de Lille is based on a genetic approach. The stereo matching problem is turned into an optimization task where the function representing the constraints of the solution is to be minimized [43].

Furthermore, to decrease the intrinsic complexity of stereo vision, some domain specific constraints are generally adopted.

In the GOLD system, the removal of the perspective effect from stereo images allows to obtain two images that differ only where the initial assumption of a flat road is not valid, thus detecting the free space in front of the vehicle [32].

Analogously, the University of California research unit developed an algorithm that remaps the left image using the point of view of the right image, thus detecting disparities in correspondence to the obstacles. A Kalman filter is then used to track obstacles [44].

Table 2 compares the strong and weak points of the different approaches to the obstacle detection problem.

Table 2: Comparison of the different approaches to obstacle detection

	Pros	Cons
Analysis of a single image	Simple algorithmic solutions, fast processing, does not suffer from vehicle movement	Loss of info about scene's depth unless specific assumptions are made, not successful when obstacles do not match the model
Optical flow	Detection of generic objects, allows computation of ego-motion and obstacles' relative speed	Computationally complex, sensitive to vehicle movements and drifts in calibration, fails when both vehicle and obstacle have small or null speed
Stereo vision	Detection of generic objects, allows 3D reconstruction	Computationally complex (still domain specific constraints may reduce complexity), sensitive to vehicle movements and drifts in calibration

As mentioned above, a great deal of different techniques have been proposed in the literature and tested on a number of vehicle prototypes in order to solve the Road Following problem, but only few of them provide an integrated solution (e.g. Lane Detection and Obstacle Detection) which, obviously, leads to both an improved quality of the results and to a faster and more efficient processing.

The research group of the Istituto Elettrotecnico Nazionale “G. Ferraris" began with limiting the processing to the image portion that is assumed to represent the road, thus relying on the previously discussed Lane Detection module. This area of the image was analyzed and borders that could represent a potential vehicle were looked for and examined [31].

Moreover, there are situations where the combination of Lane Detection and Obstacle Detection is mandatory. This is true, for example, for those systems which focus on the analysis of the vehicle’s rear view, with the aim of increasing driver and passengers safety. Without a knowledge of lane structure, in fact, it would be very difficult (if not impossible) to estimate the exact positions of the following vehicles.

Some researchers focus on the implementation of electronic rear-view mirrors, which assist the driver in analyzing what occurs on the road behind him or her. As an example, a system developed at the University of Amsterdam uses a single camera to derive the real-world motion of the vehicles behind the car. Information which can be drawn includes Time to Contact (to avoid bumper-to-bumper crashes) and lane shifts (useful during overtakings) [45]. The system implemented by Daimler Chrysler AG Research Institute and the Universität der Magdeburg, instead, detects vehicles in the rear view of the host car by means of two cameras, thus exploiting stereovision. Using the steering angle and the detected obstacles, the trajectory of the car can be properly reconstructed and used as a lane change assistant [46].

3.3 Pedestrian Detection

Vision-based pedestrian detection in outdoor scenes is still an open challenge. People dress in very different colors that sometimes blend with the background, they wear hats or carry bags, and stand, walk and change direction unpredictably. The background is various, containing buildings, moving or parked cars, cycles, street signs, signals etc. Moreover, sudden changes of background are inevitable in vision systems mounted on a moving vehicle.

Many different approaches have been developed to address this complexity. Pattern analysis, stereo vision, shape detection and tracking have been fused in more than one combination. Only few of these systems have already proved their efficacy in applications to intelligent vehicles. Nonetheless, all the principal trends in research will be discussed below to give a broad view on this highly developing field.

The most common approach to pedestrian detection consists in two conceptual steps. First, the image is segmented into foreground and background regions. Then, a second step determines if a foreground region is a pedestrian or not. Most of the work concerning the detection of human shape in cluttered scenes is due to studies on automatic surveillance systems. The fundamental assumption of this research field is a fixed or slow moving camera. In such a situation, a common approach to the detection of regions of interest is to subtract each single frame form a reference frame or an intensity model of the empty field of view built at initialization.

Although very effective, this premise does not fit the requirements of a system for automotive applications, where the background is continuously changing and no modeling is reasonably achievable. Alternative ways of segmentation have been employed, in particular those involving the analysis of more than one image, such as

- the analysis of *motion* and
- the processing of *stereo images*.

Motion is a common cue to detect interesting regions in a scene. It heavily uses temporal information and has proved to be quite reliable if one wants only to find a moving object, and not its precise velocity. Unfortunately, it does not detect standing pedestrians and needs the analysis of a sequence of a few frames before giving a response.

A few works use motion detection with optical flow as a means of segmentation. The basic idea is to detect blobs with a given shape or a common feature, like color, that have similar values of optical flow, and track their movement in subsequent frames. A group of the University of Rochester [47] analyzes the scene with a discrete XYT cube, where they assign to each region its average optical flow. An XYT cube is a representation of a sequence of T frames, each divided in XxY areas in its two spatial dimensions. In this system, four divisions are used in each spatial dimension, and six along the temporal dimension, resulting in a feature vector of size 96 containing the average optical flow of each region. They then employ Fourier analysis to classify these values. Their method has been applied to the monitoring of some repetitive human activities with a stationary camera, like walking or running on a on gymnastic rolling belt. Like other methods based on optical flow, a good cancellation of ego motion is critical in applications with a moving camera.

Some groups suggest alternative ways for motion detection. A system devised for surveillance applications by the Queen Mary and Westfield College, London, [48] uses a zero-crossing detection algorithm using the convolution of a spatio-temporal Gaussian with the history over the values of a pixel in the past 6 frames. This gave good results, with the extension of a second-order Kalman filter that copes with occlusions. An original work by *Cutler and Davis* of the University of Maryland [49] uses a subtraction between an image at time *t*, and a version of the same image stabilized with respect to image at instant *t*-τ. This operation, followed by an appropriate thresholding, gives a map of the pixels representing moving objects. Their method is then based on a two-step approach where recognition is made through analysis of the correlation of two frames taken with a delay of τ. The authors report good performance both with stationary cameras and moving cameras, provided that the background is homogeneous to some extent.

A different approach to the segmentation problem is range thresholding based on *stereo analysis.* In their Bus Driver Assistance project, *Zhao and Thorpe* of the Carnegie Mellon University [50] use range as a means of object segmentation. Range stereo systems rely on some kind of correlation between left and right images of the same scene and are strongly affected by noise, above all at longer ranges. The authors report problems in the detection procedures due to the fact that each segmented region does not always correspond to a single object. An hypothesis and verification procedure is then necessary to split or group segmented regions, which are then passed to a pedestrian recognition module.

In surveillance applications, stereo analysis is sometimes used as a cue to build a disparity map of the background for use with background subtraction. This is what happens in the system realized by SRI International using their integrated stereo cameras called Small Vision System (SVS) [51]. Segmented objects are therefore organized in pyramids to compensate for scale differences. This system shows low sensibility to distracting elements like shadows, lighting changes, occluding objects or camera dynamics. Moreover, foreground regions may be separated even if they are at the same distance as some background features.

Some systems substitute the segmentation step with a focus-of-attention approach, where salient regions in opportune feature maps are interpreted as candidates for pedestrians. In the GOLD system [52], vertical symmetries are associated with candidates for standing pedestrians, both moving and stationary. Further information derives from symmetry maps of horizontal edges and of their number per column. Then a bounding box encloses interesting regions for a separated recognition step. A more complicated system has been developed at the Ruhr-Universität Bochum [53]. The focus of attention is directed by a composition of a map of the local image entropy, a model-matching module with the shape of a Λ representing human legs, and finally inverse perspective mapping (binocular vision) for the short distance field. This information is combined in a temporal dynamic activation field (DAF) that efficiently allocates computational resources in the following recognition and tracking step.

For what concerns the recognition phase, two main trends are pursued in recent research:

- detection of the typical periodicity of *human gait* in the movement of foreground regions, and
- *shape analysis* of foreground regions.

Methods based gait recognition show a higher robustness, but they require the analysis of multiple frames and easily apply only to pedestrians crossing the street in the path of the vehicle, where the alternating movement of legs is more evident. An important drawback of this family of systems is their inability to correctly classify still persons as pedestrians. On the other hand, shape-based approaches are more sensible to false positives and thus they need a good detection phase, but they correctly recognize even stationary people.

Periodicity of the *human gait* is often recognized with traditional methods like the Fourier transform. Some systems perform a frequency analysis of the changes of candidate patterns over time, and then select those that show the frequency spectrum characteristic of human gait. As an example *Cutler and Davis* [49] use a Short-Time Fourier Transform with a Hanning windowing function to analyze the signals obtained by correlation of the pattern of detected objects.

In one of the studies for the development of the UTA [54], an Adaptable Time Delay Neural Network (ATDNN) algorithm is considered. After a first stereo-based segmentation that detects and extracts image regions containing legs of pedestrians, the ATDNN performs a local spatio-temporal processing to detect the typical pattern of the movement. This way, the gait patterns of a pedestrian in a complete gait cycle are learned by the network.

In the algorithm by the Ruhr-Universität Bochum [53], the torso of a candidate pedestrian is tracked so that the lower part of the region can be analyzed to reveal the relative motion of legs. A rough model of two legs consisting of two rod-like pieces each, jointed at the knees, is juxtaposed on the image area below the tracked torso. The periodic movement detected is then correlated to an experimental curve derived from the statistical average of human gait periods. High peaks of the correlation function indicate the presence of a person.

Basic *shape analysis* methods consist in matching a significant and simple shape onto candidate foreground regions. Some systems, like GOLD [52] or the one by SRI International [51], employ a Ω model for the head and shoulders. This approach is very sensible to scale variation, so multiple models of different scales are needed. In the two systems above, three and five different models are used, from coarse to fine resolution, according to the estimated distance of a subject. A group at The Robotics Institute of the Carnegie Mellon University [55] uses a skeletonization procedure to characterize the shape of a foreground object previously detected. For each object, they calculate first the centroid of the area and then the distances from the centroid to each border points. Local maxima of the distance function are taken as the external points of the skeleton. The authors suggest that the relative position of centroid and external points, and their rigidity, may be applied to recognition of different types of targets. For what concerns humans, they further confirm the analysis with gait detection.

Another algorithm developed within the UTA project at DaimlerChrysler [56] presents a two-step approach where both phases rely on shape and pattern analysis. The detection step is based on a search of the image with a numerous set of silhouettes of the human body using a distance transform of the edge image. The silhouettes are organized hierarchically with a coarse-to-fine approach, so that generic forms are tried first, and similar and more detailed shapes afterwards. The validation step is accomplished by a radial-basis-function classifier trained with rectangular regions containing pedestrians which were previously selected by a human operator. In a work by the University of Maryland, the system was further improved to perform tracking [57]. A statistical shape model of a pedestrian is first built and then approximated by a Linear Point Distribution Model. The tracking of this model over the image sequence is accomplished with a quasi-random sampling method, based on a zero-order motion model with large process noise high enough to account for the greatest expected change in shape and motion. The authors report a high rate of success and the ability of the tracker to quickly recover from failures.

More systems employ pattern recognition with classifiers to accomplish the recognition step. Sometimes the original imaged is processed before the application of the classifier. For example, *Zhao and Thorpe* [50] propose a three-layer feed forward network processing the intensity gradient image rather than the original image.

Table 3 summarizes the strong and weak points of the different approaches to the pedestrian detection problem.

Table 3: Pros and Cons of the most typical assumptions in lane detection

	Pros	Cons
Segmentation with motion	Good reliability if background is still or changing slowly, gives strong information for gait detection	Cancellation of ego-motion is needed, needs analysis of a sequence of images
Segmentation with stereo	Low sensibility to lighting changes or occlusions, feasible even with changing backgrounds	Does not segment single objects (hypothesis and verification procedure is needed), unreliable at longer ranges
Focus of attention	Analysis carried out on a single-image basis, fast processing	Features must be carefully selected, does not really segment an object from the background
Recognition of human gait	Gait frequency is highly recognizable, high reliability	Does not recognize standing pedestrians, limited to pedestrians crossing the path of the vehicle, needs analysis of a sequence of images
Recognition of human shape	More versatile, feasible on a single image, fast processing	Problems at multiple scales, human shape has a high variability, difficult if segmentation is not precise

The system devised at the AI Lab of the M.I.T. [58] for automotive applications fuses the detection and validation steps into a single one. The image is initially transformed with Haar wavelets and then scanned to detect the pattern associated with a human person. The human pattern is learned, and subsequently recognized, through statistical reasoning with a support vector machine – a technique to train classifiers which is capable of learning in sparse, high-dimensional spaces with very few examples. The system uses multiple classifiers for arms, head and legs, in a hierarchical organization, in order to cope with occlusions. In [59], the system was adapted to consider temporal information in the form of a joint analysis of five sequential frames.

4. Architectural Issues

In the early years of ITS applications a great deal of custom solutions were proposed, based on ad hoc, special-purpose hardware. This recurrent choice was motivated by the fact that the hardware available on the market at a reasonably low cost was not powerful enough to provide real-time image processing capabilities. As an example, the researchers of the Universität der Bundeswehr developed their own system architecture: several special-purpose boards were included in the Transputer-based architecture of the VITA vehicle [60].

Others developed or acquired ad hoc processing engines based on SIMD computational paradigms to exploit the spatial parallelism of images. Among them, the cases of the 16k Mas-Par MP-2 installed on the experimental vehicle NavLab I [61], [62] at the Carnegie Mellon University (CMU) and the massively parallel architecture PAPRICA [63] jointly developed by the University of Parma and the Politecnico di Torino and tested on the MOB-LAB vehicle.

Besides selecting the proper sensors and developing specific algorithms, a large percentage of this first research stage was therefore dedicated to the design, implementation, and test of new hardware platforms. In fact, when a new computer architecture is built, not only the hardware and architectural aspects -such as instruction set, I/O interconnections, or computational paradigm- need to be considered, but software issues as well. Low-level basic libraries must be developed and tested along with specific tools for code generation, optimization and debugging.

In the last few years, the technological evolution led to a change: almost all research groups are shifting towards the use of of-the-shelf components for their systems. In fact, commercial hardware has nowadays reached a low price/performance ratio. As an example, both the NavLab 5 vehicle from CMU and the ARGO vehicle from the University of Parma are presently driven by systems based on general-purpose processors. Thanks to the current availability of fast internetworking facilities, even some MIMD solutions are being explored, composed of a rather small number of powerful, independent processors, as in the case of the VaMoRs-P vehicle of the Universität der Bundeswehr on which the Transputer processing system has now been partly replaced by a cluster of three PCs (dual Pentium II) connected via a Fast Ethernet-based network [20].

Current trends, however, are moving towards a mixed architecture, in which a powerful general-purpose processor is aided by specific hardware such as boards and chips implementing optical flow computation, pattern-matching, convolution, and morphological filters. Moreover, some SIMD capabilities are now being transferred into the instruction set of the last-generation CPUs, which has been tailored to exploit the parallelism intrinsic to the processing of visual and audio (multimedia) data. The MMX extensions of the Intel Pentium processor, for instance, are exploited by the GOLD system which acts as the automatic driver of the ARGO vehicle to boost up performance.

In conclusion, it is important to emphasize that, although the new generation systems are all based on commercial hardware, the development of custom hardware has not lost significance, but is gaining a renewed interest for the production of embedded systems. Once a hardware and software prototype has been built and extensively tested, its functionalities have to be integrated in a fully optimized and engineered embedded system before marketing. It is in this stage of the project that the development of ad hoc custom hardware still plays a fundamental role and its costs are justified through a large scale market.

5. Perspectives on Intelligent Vehicles

The promising results obtained in the first stages of research on intelligent vehicles demonstrate that a full automation of traffic (at least on motorways or sufficiently structured roads) is technically feasible.

Nevertheless, besides technical problems some issues must be carefully considered in the design of these systems such as legal aspects related to the responsibility in case of faults and incorrect behavior of the system, and the impact of automatic driving on human passengers. User acceptance in particular will play a critical role in how intelligent vehicles will look and perform and the system interface will have a strong influence on how a user will view and understand the functionality of the system.

Therefore, a long period of exhaustive tests and refinement must precede the availability of these systems on the general market, and a fully automated highway system with intelligent vehicles driving and exchanging information is not expected for a couple of decades.

For the time being, complete automation will be restricted to special infrastructures such as industrial applications or public transportation. Then, automatic vehicular technology will be gradually extended to other key transportation areas such as goods shipping, for example on expensive trucks, where the cost of an autopilot is negligible with respect to the cost of the vehicle itself and the service it provides. Finally, once technology has stabilized and the most promising solution and best algorithms fixed, a massive integration and a widespread use of such systems will take place in private vehicles, but this will not happen for another two or more decades.

References

[1] C. Little, “The Intelligent Vehicle Initiative: Advancing ‘Human-Centered’ Smart Vehicles,” *Public Roads Magazin*e, vol. 61, no. 2, Sept./Oct. 1997, pp. 18–25.

[2] M. Cellario, “Human-Centered Intelligent Vehicles: Toward Multimodal Interface Integration,” *IEEE Intelligent Systems*, pp. 78–81, Jul./Aug. 2001.

[3] D. Bishop, "Vehicle-Highway Automation Activities in the United States, in *Proc. Intl. AHS Workshop*, U. S. Department of Transportation, 1997.

[4] H. Tokuyama, "Asia-Pacific Projects Status and Plans", in *Proc. Intl. AHS Workshop. U. S. Department of Transportation*, 1997.

[5] M.C. Hulse et al., *Development of Human Factors Guidelines for Advanced Traveler Information Systems and Commercial Vehicle Operations: Identification of the Strengths and Weaknesses of Alternative Information Display Format*s, tech. report FHWA-RD-96-142, Federal Highway Administration, Washington, D.C., 1998.

[6] U. Seger, H. G. Graf, and M. E. Landgraf, "Vision Assistance in Scenes with Extreme Contrast", *IEEE Micro*, pp. 50-56, Jan.-Feb. 1993.

[7] C. G. Sodini and S. J. Decker, "A 256 x 256 CMOS Brightness Adaptive Imaging Array with Column-Parallel Digital Output", in Proc. IEEE IV, 1998, pp. 347-352.

[8] M. Mizuno, K. Yamada, T. Nakano, and S. Yamamoto, "Robustness of Lane Mark Detection with Wide Dynamic Range Vision Sensor", in *Proc. IEEE IV*, 1995, pp. 171-176.

[9] T. M. Jochem, D. A. Pomerleau, and C. E. Thorpe, "MANIAC: A Next Generation Neurally Based Autonomous Road Follower", in *Proc. 3rd Intl. Conf. on Intelligent Autonomous Systems*, 1993.

[10] J. D. Crisman and C. E. Thorpe, "UNSCARF, A Color Vision System for the Detection of Unstructured Roads", in *Proc. IEEE Intl. Conf. on Robotics and Automation*, 1991, pp. 2496-2501.

[11] K. I. Kim, S. Y. Oh, S. W. Kim, H. Jeong, C. N. Lee, B. S. Kim, and C. S. Kim, "An Autonomous Land Vehicle PRV II: Progresses and Performance Enhancement", in *Proc. IEEE IV*, 1995, pp. 264-269.

[12] A. Takahashi, Y. Ninomiya, M. Ohta, and K. Tange, "A Robust Lane Detection using Real-time Voting Processor", in *Proc. IEEE ITS*, 1999, pp. 577-580.

[13] S. L. Michael Beuvais, Chris Kreucher, "Building World Model for Mobile Platforms using Heterogeneous Sensors Fusion and Temporal Analysis", in *Proc. IEEE ITS*, 1997, p. 101.

[14] A. Coda, P. C. Antonello, and B. Peters, "Technical and Human Factor Aspects of Automatic Vehicle Control in Emergency Situations", in *Proc. IEEE ITS*, 1997.

[15] D. A. Pomerleau and T. Jochem, "Rapidly Adapting Machine Vision for Automated Vehicle Steering", *IEEE Expert*, 11(2), Apr. 1996.

[16] J. P. Gonzàlez and Ü. Özgüner, "Lane Detection Using Histogram-Based Segmentation and Decision Trees", in *Proc. IEEE ITS*, 2000, pp. 346-351.

[17] P. Charbonnier, P. Nicolle, Y. Guillard, and J. Charrier, "Road Boundaries Detection using Color Saturation", in *Proc. 9th European Signal Processing Conf.* '98, Sept. 1998.

[18] R. Chapuis, R. Aufrère, F.Chausse, and J. Alizon, "Road Sides Recognition under Unfriendly Lighting Conditions", in *Proc. IEEE IV*, 2001, pp. 13-18.

[19] J. Goldbeck, D. Graeder, B. Huertgen, S. Ernst, and F. Wilms, "Lane Following Combining Vision and DGPS", in *Proc. IEEE IV*, 1998, pp. 445-450.

[20] M. Lützeler and E. D. Dickmanns, "Road Recognition with MarVEye", in *Proc. IEEE IV*, 1998, pp. 341-346.

[21] U. Franke, D. Gavrila, S. Görzig, F. Lindner, F. Paetzold, and C. Wöhler, "Autonomous Driving Goes Downtown", in *Proc. IEEE IV*, 1998, pp. 40-48.

[22] J. Goldbeck and B. Huertgen, "Lane Detection and Tracking by Video Sensors", in *Proc. ITS*, 1999, pp. 74-79.

[23] K. A. Redmill, S. Upadhya, A. Krishnamurthy, and Ü. Özgüner, "A Lane Tracking System for Intelligent Vehicle Applications", in *Proc. IEEE ITS*, 2001, pp. 275-281.

[24] K. A. Redmill, "A Simple Vision System for Lane Keeping", in *Proc. IEEE ITS*, 1997.

[25] F. Chausse, R. Aufrère, and R. Chapuis, "Vision Based Vehicle Trajectory Supervision", in *Proc. IEEE ITS*, 2000, pp. 143-148.

[26] J. Goldbeck and B. Huertgen, “Lane Detection and Tracking by Video Sensors”, in *Proc. ITS*, 1999, pp. 74-79.

[27] R. Risack, P. Klausmann, W. Kr. uger, and W. Enkelmann, “Robust Lane Recognition Embedded in a Real-Time Driver Assistance System”, in *Proc. IEEE IV*, 1998, pp. 35-40.

[28] S. M. Wong and M. Xie, “Lane Geometry Detection for the Guidance of Smart Vehicle”, in *IEEE ITS*, 1999, pp. 925-928.

[29] X. Youchun, W. Rongben, and J. Shouwen, “A Vision Navigation Algorithm Based on Linear Lane Model”, in *Proc. IEEE IV*, 2000, pp. 240-245.

[30] A. Broggi and S. Bertè, “Vision-Based Road Detection in Automotive Systems: a Real-Time Expectation-Driven Approach”, *Journal of Artificial Intelligence Research*, 3:325-348, Dec. 1995.

[31] S. Denasi, C. Lanzone, P. Martinese, G. Pettiti, G. Quaglia, and L. Viglione, “Real-time system for road following and obstacle detection”, in *Proc. SPIE on Machine Vision Applications, Architectures, and Systems Integration III*, 1994, pp. 70-79.

[32] M. Bertozzi and A. Broggi, “GOLD: a Parallel Real-Time Stereo Vision System for Generic Obstacle and Lane Detection”, *IEEE Trans. on Image Processing*, 7(1):62-81, Jan. 1998.

[33] R. Aufrère, R. Chapuis, and F. Chausse, “A Fast and Robust Vision Based Road Following Algorithm”, in *Proc. IEEE IV*, 2000, 192-197.

[34] A. Broggi, M. Bertozzi, A. Fascioli, and G. Conte, “Automatic Vehicle Guidance: the Experience of the ARGO Vehicle”, *World Scientific*, 1999.

[35] S. Kyo, T. Koga, K. Sakurai, and S. Okazaki, “A Robust Vehicle Detecting and Tracking System for Wet Weather Conditions using the IMAP-VISION Image Processing Board”, in *Proc. IEEE ITS*, 1999, pp. 423-428.

[36] S. Denasi and G. Quaglia, “Obstacle Detection Using a Deformable Model of Vehicles”, in *Proc. IEEE IV*, 2001, pp. 145-150.

[37] W. Kruger, W. Enkelmann, and S. Rossle, “Real-Time Estimation and Tracking of Optical Flow Vectors for Obstacle Detection”, in *Proc. IEEE IV*, 1995, pp. 304-309.

[38] S. M. Smith and J. M. Brady, “ASSET-2: Real-time motion segmentation and shape tracking”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(8):814-829, 1995.

[39] Z. Hu and K. Uchimura, “Tracking Cycle: A New Concept for Simultaneously Tracking of Multiple Moving Objects in a Typical Traffic Scene”, in *Proc. IEEE IV*, 2000, pp. 233-239.

[40] F. Marmoiton, F. Collange, and J. P. Dèrutin, “Location and Relative Speed Estimation of Vehicles by Monocular Vision”, in *Proc. IEEE IV*, 2000, pp. 227-232.

[41] H. Hattori, “Stereo for 2D Visual Navigation”, in *Proc. IEEE IV*, 2000, pp. 31-38.

[42] M.Hariyama, T. Takeuchi, and M. Kameyama, “Reliable Stereo Matching for Higly-Safe Intelligent Vehicles and Its VLSI Implementation”, in *Proc. IEEE IV*, 2000, pp. 128-132.

[43] Y. Ruichek, H. Issa, and J. Postaire, “Genetic Approach for Obstacle Detection using Linear Stereo Vision”, in *Proc. IEEE IV*, 2000, pp. 261-266.

[44] D. Koller, J. Malik, Q.-T. Luong, and J. Weber, “An integrated stereo-based approach to automatic vehicle guidance”, in *Proc. 5th Intl. Conf. on Computer Vision*, 1995, pp. 12-20.

[45] M. B. van Leeuwen and F. C. A. Groen, “Motion Estimation with a Mobile Camera for Traffic Applications”, in *Proc. IEEE IV*, 2000, pp. 58-63.

[46] C. Knoeppel, A. Schanz, and B. Michaelis, “Robust Vehicle Detection at Large Distance Using Low Resolution Cameras”, in *Proc. IEEE IV*, 2000, pp. 267-272.

[47] R. Polana and R. C. Nelson, “Detection and Recognition of Periodic, Non-rigid Motion”, *Int. J. Comp. Vis.*, vol. 23(3), pp. 261-282, 1997.

[48] S. J. McKenna and S. Gong, “Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition”, in *Int. Conf Audio and Video Authentication*, 1997, pp. 177-184.

[49] R. Cutler and L. S. Davis, “Robust Real-time Periodic Motion Detection, Analysis and Applications”, *IEEE Trans. Patt. An. Mach. Int.*, vol. 22(8), pp. 781-796, 2000.

[50] L. Zhao and C. Thorpe, “Stereo and Neural Network Based Pedestrian Detection”, *IEEE Trans. Int. Transp. Sys.*, 1(3), pp. 148-154, 1999.

[51] D. Beymer and K. Konolige, “Real-time Tracking of Multiple People Using Continuous Detection”, in *Proc. Int. Conf. Comp. Vis.*, 1999.

[52] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi, “Shape-based Pedestrian Detection”, in *Proc. IEEE IV*, 2000, pp. 215-220.

[53] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, “Walking Pedestrian Recognition”, *IEEE Trans Int Transp. Sys*, vol. 1(3), pp. 155-163, 2000.

[54] C. Wöhler, U. Kressler, and J. K. Anlauf, “Pedestrian Recognition by Classification of Image Sequences. Global Approaches vs Local Spatio-temporal Processing”, in *Proc. IEEE Int. Conf. Patt. Rec.*, 2000.

[55] H. Fujiyoshi and A. Lipton, “Real-time Human Motion Analysis by Image Skeletonisation”, in *Proc. IEEE WACV’98*, 1998, pp. 15-21.

[56] D. M. Gavrila, “Pedestrian Detection from a Moving Vehicle”, in *Proc. Eur. Conf. Comp. Vis.*, 2000, pp. 37-49.

[57] V. Philomin, R. Duraiswami, and L. Davis, "Pedestrian Tracking from a Moving Vehicle", in *Proc. IEEE IV*, 2000, pp. 350-355.

[58] A. Mohan, C. Papageorgiou, and T. Poggio, “Example-based Object Detection in Images by Components”, *IEEE Trans. Patt. An. Mach. Int.*, vol. 23(4), pp. 349-361, 2001.

[59] C. Papageorgiou and T. Poggio, “A Pattern Classification Approach to Dynamical Object Detection”, in *Int. Conf. Comp. Vis.*, 1999, pp. 1223-1228.

[60] E. D. Dickmanns, “Expectation-Based Multi-Focal Vision for Vehicle Guidance, in *Proc. 8th European Signal Processing Conf.*, 1995, pp. 1023-1026.

[61] T. M. Jochem and S. Baluja, “A Massively Parallel Road Follower”, in M. A. Bayoumi, L. S. Davis, and K. P. Valavanis, editors, *Proc. IEEE Computer Architectures for Machine Perception*, 1998, pp. 2-12.

[62] T. M. Jochem and S. Baluja, “Massively Parallel, Adaptive, Color Image Processing for Autonomous Road Following”, in H. Kitano, editor, *Massively Parallel Artificial Intelligence*, AIII Publishers in cooperation with MIT Press, 1993.

[63] A. Broggi, G. Conte, F. Gregoretti, C. Sansoè, and L. M. Reyneri, “The Evolution of the PAPRICA System”, *Integrated Computer-Aided Engineering Journal* - Special Issue on Massively Parallel Computing, 4(2):114-136, 1997.

Vision-based Pedestrian Detection: will Ants Help?

M. Bertozzi, A. Broggi, A. Fascioli, and P. Lombardi

Abstract— **This work presents the vision-based system for detecting pedestrians in road environments implemented on the ARGO prototype vehicle developed by the University of Parma. The system is aimed at the localization of pedestrians in various poses, positions and clothing, and is not limited to moving people.**

Initially, attentive vision techniques relying on the search for specific characteristics of pedestrians such as vertical symmetry and strong presence of edges, allow to select interesting regions likely to contain pedestrians. Then, such candidates areas are validated verifying the actual presence of pedestrians by means of an shape detection technique based on the application of autonomous agents.

Keywords— **pedestrian detection, machine vision, autonomous agents, symmetry, ants.**

I. Introduction

The detection of pedestrians is a mandatory requirement for future driving assistance systems, since having the capability of avoiding crushes with pedestrians is essential for effectively aiding the driver in urban environments.

This work presents the vision-based system for detecting pedestrians implemented on the ARGO vehicle. ARGO is an experimental autonomous vehicle developed by the University of Parma, equipped with a vision system, a processing engine, and automatic steering capabilities [1]. The main target of the ARGO Project is the development of an active safety system which can also act as an automatic pilot for a standard road vehicle.

Vision-based pedestrian detection in outdoor scenes is a challenging task even in the case of a stationary camera. In fact, pedestrians usually wear different clothes with various colors that, sometimes, are barely distinguishable from the background (this is particularly true when processing black and white images). Moreover, pedestrians can wear or carry items like hats, bags, umbrellas, and many others, which give a broad variability to their shape.

When the vision system is installed on-board of a moving vehicle additional problems must be faced, since the observer's ego-motion entails additional motion in the background and changes in the illumination conditions. In addition, since Pedestrian Detection is more likely to be of use in a urban environment, also the presence of a complex background (including buildings, moving or parked cars, cycles, road signs, signals, ...) must be taken into account.

Widely used approaches for addressing vision-based Pedestrian Detection are: the search of specific patterns or textures [2], stereo vision [3, 4], shape detection [5, 6, 7], motion detection [8, 9, 10], neural networks [11, 12]. The great part of the research groups use a combination of two or more of these approaches [13, 14, 15]. Anyway, only a few of these systems have already proved their efficacy in applications for intelligent vehicles.

In this work the strong vertical symmetry of the human shape is exploited to determine specific regions of interest which are likely to contain pedestrians. Subsequently, an agent-based algorithm is used to verify the presence of pedestrians in those areas through an innovative application of a shape detection technique. This method allows the identification of pedestrians in various poses, positions and clothing, and is not limited to walking people.

This paper is organized as follows. Section 2 introduces the attentive algorithm aimed at the determination of the areas of interests, and shows its results. Section 3 describes the agent-based algorithm used for the validation of the candidates and presents the preliminary results of this shape detection technique. Section 4 ends the paper with some final remarks.

II. Attentive vision

As a first processing step, attentive vision techniques are applied to concentrate the analysis on specific regions of interest only. In fact, the aim of the low-level part of the processing is the focusing on potential candidate areas to be further examined at a higher-level stage in the following steps.

The areas considered as candidate are rectangular bounding boxes which:

- have a size in pixels deriving from the knowledge of the intrinsic parameters of the vision system (angular aperture and resolution); in other words, once defined the size and distance of a pedestrian in the 3D world (e. g. 1.8 $m \times 0.6$ m at 20 m), simple perspective considerations give the size in pixel of its projection in the image (see figure 1.a);
- enclose a portion of the image which exhibits the low-level features that characterize the presence of a pedestrian, i. e. a strong vertical symmetry and a high density of vertical edges.

These bounding boxes will be then checked against a human shape model, taking into account the contour of the object they enclose, in order to be validated.

The search for candidates would require an exhaustive search in the whole image (see figure 1.b). However, the knowledge of the system's extrinsic parameters, together with a flat scene assumption (see figure 1.c), is exploited to limit the analysis to a stripe of the image (hereinafter referred to as *search area*). The displacement of this stripe depends on the pedestrian's distance, while its height is related to the pedestrian's height (see figure 1.d). Since by definition a pedestrian is a human shaped road participant, the flat world assumption becomes an assumption on the road slope, which is anyway a loose hypothesis in a road environment, particularly in the area immediately ahead of the vehicle. Besides the obvious advantage of avoiding false detections in wrong areas, the processing of the search

This work was supported by the European Research Office of the U. S. Army under contract number N68171-01-M-5857

M. Bertozzi, A. Broggi, and A. Fascioli are with the Dip. di Ingegneria dell'Informazione, Università di Parma, ITALY. E-mail: {bertozzi,broggi,fascioli}@ce.unipr.it.

P. Lombardi is with the Dip. di Informatica e Sistemistica, Università di Pavia, ITALY. E-mail: lombardi@vision.unipv.it

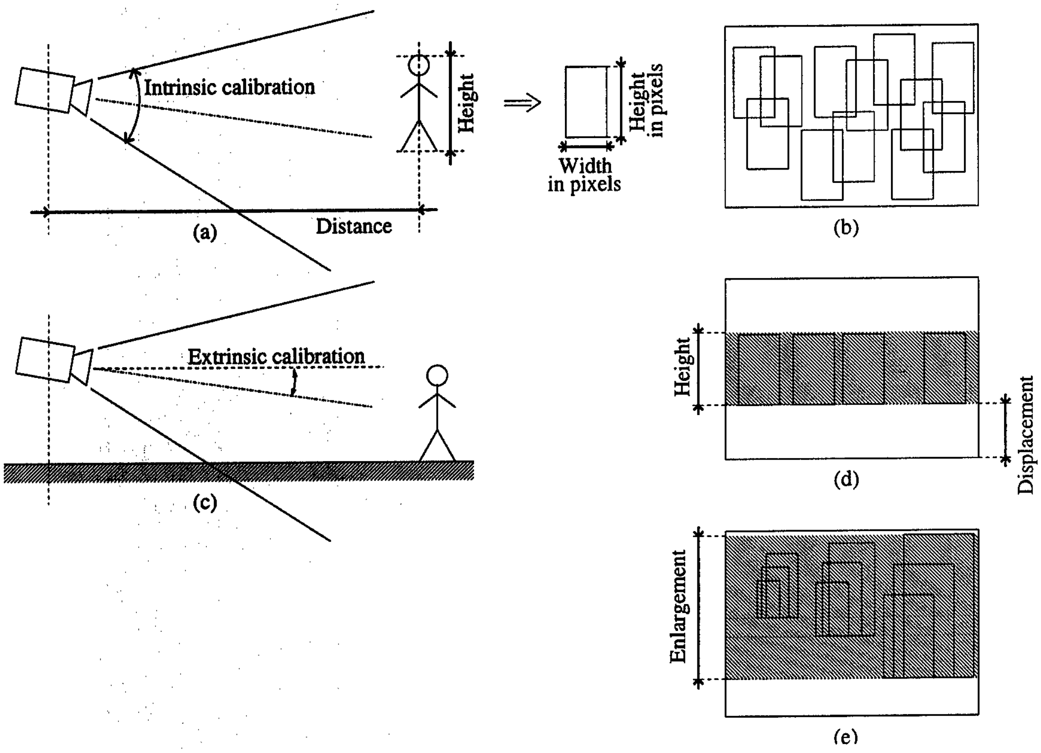


Fig. 1 *(a)* Computation of the bounding box size given the intrinsic parameters and the size and distance of a pedestrian; *(b)* exhaustive search for candidates in the whole image; *(c)* the search area can be limited to a stripe given the extrinsic parameters and a flat scene assumption; *(d)* the displacement and height of the stripe depend on the pedestrian distance and height, respectively; *(e)* the search area is enlarged to explore a range of distances and heights.

area only reduces the computational time. Indeed, the analysis cannot be limited to a fixed size and distance of the target and a given range for each parameter is in fact explored (e. g. $1.6 \div 2.0\ m \times 0.5 \div 0.7\ m$ at $10 \div 30\ m$). The introduction of these ranges generates two further degrees of freedom in the size and position of the bounding boxes. In other words, the search area is enlarged to accommodate all possible combinations of height, width, and distance (see figure 1.e).

The analysis proceeds in this way: the columns of the image are considered as possible symmetry axes for bounding boxes. For each symmetry axis different bounding boxes are evaluated scanning a specific range of distances from the camera (the distance determines the position of the bounding box base) and a reasonable range of heights and widths for a pedestrian (the corresponding bounding box size can be computed through the calibration).

However, not all the possible symmetry axes are considered: since edges are chosen as discriminant in most of the following analysis, a pre-attentive filter is applied, aimed at the selection of the areas with a high density of edges. In particular, for each axis the count of edge pixel is computed in a portion of the search area centered on the axis itself and as wide as the maximum bounding box width. Axes centered on regions which contain a number of edges lower than the average value are then dropped.

For each of the remaining axes the best candidate area is selected among the bounding boxes which share that symmetry axis, while having different position (base) and size (height and width). Vertical symmetry has been chosen as a main distinctive feature for pedestrians. Symmetry edge maps, e. g. the Generalized Symmetry Transform (GST) [16], have already been proposed as methods to locate interest points in the image prior to any segmentation or extraction of context-dependent information. Unfortunately, these methods are generally computationally expensive. Alternatively, two different symmetry measures are performed: one on the gray-level values (G) and one on the gradient values, considering only edges with a vertical direction (E). The selection of the best bounding box is based on maximizing a linear combination of the two symmetry measures, masked by the density of edges in the box (D), as shown in the following equation: $S = (k1 \times G + k2 \times E) \times D$. The weights $k1$ and $k2$ were experimentally determined analyzing a large number of images. Figure 2 shows the original input image, the result of a clustering operation used to improve the detection of edges, a binary image containing the vertical edges, and a number of histograms representing the maximum (i) symmetry of gray-levels, (ii) symmetry of vertical edges, and (iii) density of vertical edges among the bounding boxes examined for each axis. The histogram in figure 2.g represents the linear combination of all the above. The histograms are actually computed only for the regions of the search area with a high density of edges, however in figure 2 they are completely displayed for a better understanding. It is evident that, using the density of vertical

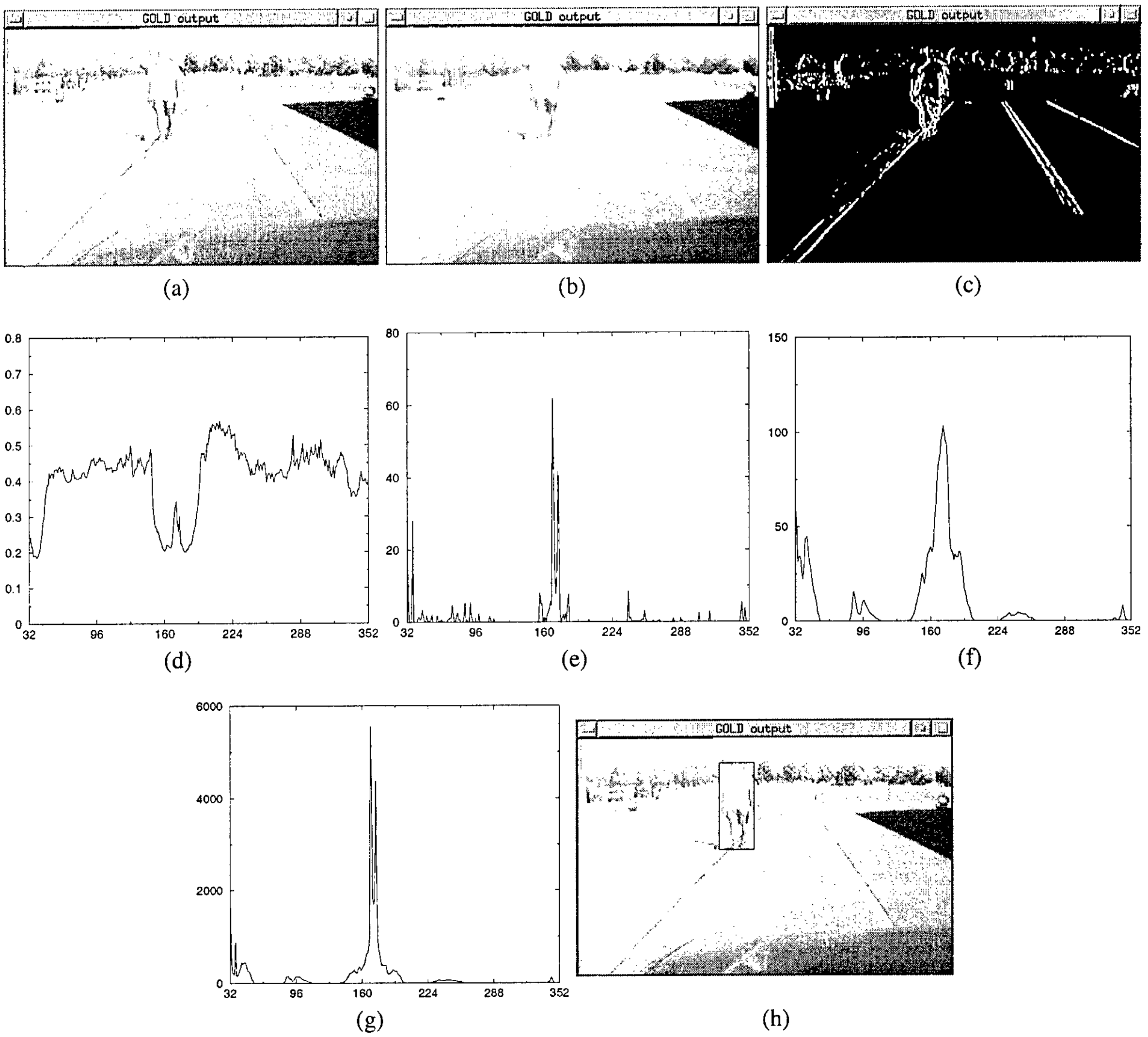


Fig. 2 Intermediate results leading to the localization of bounding boxes: *(a)* original image; *(b)* clusterized image; *(c)* vertical edges; *(d)* histogram representing grey level symmetries; *(e)* histogram representing vertical edges symmetries; *(f)* histogram representing vertical edges density; *(g)* histogram representing the overall symmetry S for the best bounding box for each column; *(h)* the resulting bounding box.

edges as a mask, interesting areas present high values for both the symmetry of gray-levels and symmetry of vertical edges. The resulting histogram is therefore thresholded and its over-threshold peaks are selected as representing candidate bounding boxes.

An adjustment of the bounding boxes' size is yet needed. In fact, when comparing the gray-level symmetry of different bounding boxes centered on the same axis, larger boxes tend to overcome smaller ones since pedestrians are generally surrounded by homogeneous areas such as concrete underneath or the sky above (this is true for other objects, too). Therefore, the bounding box which presents the maximum symmetry tends to be larger than the object it contains because it includes uniform regions. For this reason, given a peak of the overall histogram representing a selected symmetry axis, the exact height and width of the best bounding box are actually taken as those possessed by the box which maximizes a new function among the ones having the same axis. This function is computed as the product of the symmetry of vertical edges (E) and density of vertical edges (D) only. Figure 3 summarizes the overall candidate generation process.

The result of this low-level processing is a list of candidate bounding boxes which is fed to the following stage, whose task is their validation as pedestrians, based on higher-level characteristics.

A. *Results of low-level attentive vision*

The algorithm has been tested on a large number of images acquired in different situations ranging from simple uncluttered scenes to complex scenarios. As an example, figure 4 shows the result of the selection of candidate bounding boxes in three different situations. In figure 4.a a correct detection of two pedes-

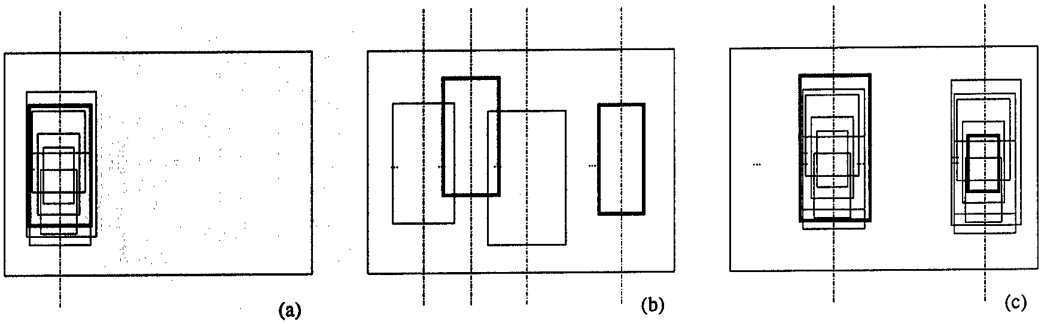


Fig. 3 *(a)* Selection of the best bounding box for each symmetry axis; *(b)* selection of the best symmetry axes; *(c)* selection of the best candidates for each selected axis by choosing the bounding box which maximizes the symmetry and density of vertical edges.

trians is displayed. Figure 4.b presents a complex scenario in which only the central pedestrian is detected; two other pedestrians are missed because (i) the first is confused with the background, and (ii) the second is only partially visible; moreover, the high symmetry of a tree has been detected. In figure 4.c the two crossing pedestrians have been localized, but other symmetrical areas are highlighted as well; the first is due to the symmetry of the area between two pedestrians, and the other is due to the presence of symmetrical road infrastructures between two trees.

Some general considerations can be drawn on the behavior of this candidate selection procedure. In situations in which pedestrians are sufficiently contrasted with respect to the background and completely visible (i. e. not hidden by other pedestrians or objects) the localization of pedestrians based on symmetry and edge density proves to be robust. Thanks to the use of vertical edges the width of the bounding boxes enclosing pedestrians is generally determined with a good precision. On the other hand, a lower accuracy is obtained for the localization of the top and bottom of the bounding box. A refinement of the bounding box height is under development.

Symmetrical objects other than pedestrians may happen to be detected as well. In order to get rid of such false positives a number of filters have been devised which rely on the analysis of the distribution of edges within the bounding box. These filters, which are still under evaluation, show promising results regarding the elimination of both artifacts (such as poles, road signs, buildings, and other road infrastructures) and symmetrical areas given by a uniform portion of the background between two foreground objects with similar lateral borders (see figure 4.c).

III. Shape Detection Using Autonomous Agents

This section describes the shape detection technique.

Different edges are selected and connected, where possible, in order to form a contour. Thanks to the way the contour is built, it will represent the shape of the pedestrian body. Matching techniques may be used in regions of the bounding box that lie in the correct position in relation to the formulated hypothesis of the pedestrian.

Essentially, the process consists in adapting a deformable coarse model to the bounding boxes. Thanks to its roughness the model is sufficiently general and can be adapted to a variety of postures. Anyway, it is limited to standing pedestrians. The model adjustment is done through an evolutionary approach with a number of independent agents acting as edge trackers. The agents explore a feature map displaying the edges contained in a given bounding box and stochastically build hypotheses of a feasible contour of a human body. The idea is taken from the Ant Colony Optimization (ACO) metaheuristic devised to solve hard combinatorial optimization problems, originally inspired by the communication behavior of real ants [17]. The system proposed here is a transposition to image analysis of one of the first ACO algorithms, the AS-*cycle*.

In nature, when ants look for food, they communicate the path and the outcome of their exploration to other ants by marking their path with a pheromone trail, its intensity depending on the distance of the food from the nest, and on its quality and quantity. Other ants are attracted by strong pheromone trails, thus the path to an abundant food source close to the nest is marked again and again until it becomes more frequented and even more attractive.

This concept can be applied to the analysis of an image by creating a colony of artificial ants that looks for an optimal combination of edge pixels that maximizes the coherency of their position according to a given model. Each ant in turn traces a solution in a solution space made up of all the possible paths connecting two pixels in a matrix. The decisional basis for each step of an ant is provided by two factors: one is a local heuristic η_i that quantifies the attractiveness of pixel i for its intrinsic characteristics; the second is the information on that pixel made available by previous attempts of other ants, in the form of a quantity of pheromone τ_i.

Artificial ants explore a world which is a matrix of pixels derived by the resampling of the edge map of the bounding box under analysis. In our experiments, the normalized world-matrix is sized 20 × 45 pixels. Each pixel i is initialized with a binary value: 1 if it contains an edge, 0 if not. This value represents its intrinsic attractiveness η_i and is the basis for the heuristic research. All the pheromone τ_i is initialized at 0. The world-matrix is visited by M ants in parallel, and the process is repeated for C cycles. At the end of each cycle, new pheromone is deposed on the trails pursued by the ants, and some of that

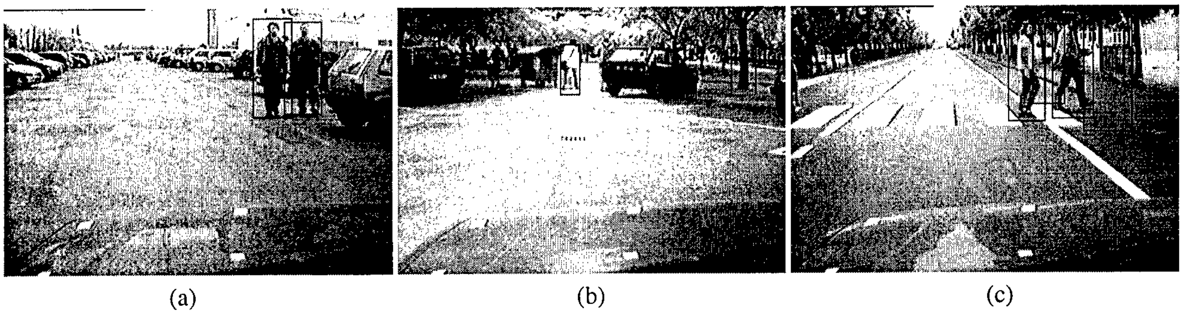


Fig. 4 Result of low-level processing in different situations: *(a)* a correct detection of two pedestrians *(b)* a complex scenario in which only the central pedestrian is detected, and the high symmetry of a tree has been detected as well; *(c)* two crossing pedestrians have been localized, but other symmetrical areas are highlighted as well.

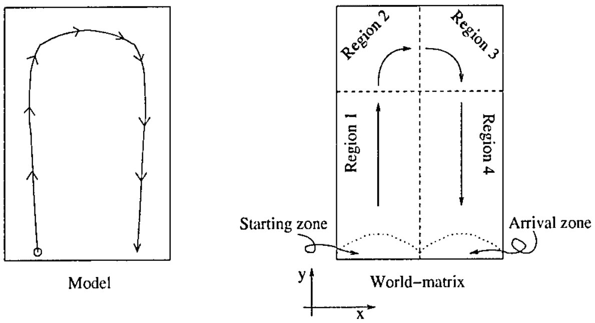


Fig. 5 Artificial ants move through the world-matrix starting from the left half of the lower border, and moving through regions 1, 2, 3 and 4 until they reach the arrival line.

accumulated evaporates. In this way, solutions built several cycles before, progressively loose their importance. On the other hand, pheromone on pixels that compose the path of frequently selected solutions grows. and eventually this information surpasses that given by the heuristic.

A crucial point to be understood is that artificial ants do not need to reach an optimal solution in the edge connection problem. Often, no real optimal solution exists even to a human inspector. The colony needs only to find a sufficiently valuable path that permits to continue the recognition, free of noisy edge pixels.

The ant system develops from a very elastic and deformable coarse model of a human body. The model is encoded in the progression rules that guide the ants through the solution space. The rules effectively restrict the whole space of possible solutions to a subspace that includes the searched shape. The system will then provide attempts to find a feasible path in this subspace, and each attempt will be evaluated by a confidence function as it is detailed in the following.

All ants start from the left half of the lower side of the world-matrix. The world is divided into four regions, as detailed in figure 5. In each region ants proceed of one step forward in the direction of one of the axis (y in regions 1 and 4, and x in regions 2 and 3), and can choose among a set of s pixels lying on the line or column in front of them. Additionally, in regions 2 and 3 ants have the option of moving vertically, thus they can follow very steep edges as well as very flat ones.

The starting point of each ant is chosen randomly among the edge pixels lying in the starting region, i.e. the left half of the lower border. If no edge pixel is present, the starting point is set on a random point belonging to the region. The choice of starting from edge pixels does not pose a hard restriction to the exploration of the solution space: the bounding box usually comprises edges in the lowest line owing to the mechanism that determines its dimensions based on the edge density. Most of the times, the edges appearing on the lower line of a well-centered bounding box correspond to the feet of the pedestrian. Each ant stops its journey when it reaches the right half of the lower border of the world-matrix.

An ant is an independent pixel-sized agent; it has a local exploratory capability, limited to the set of pixels belonging to the scanning region N as described above, and of those lying on the following line as well. Figure 6 illustrates the situation for the scanning sets for each region of the world-matrix. Each pixel under consideration is associated to a quality measure that takes into account terms pertaining to both the feature map of the edges, and the pheromone deposed by previous ants. The quality of pixel j is expressed as $q_j = \alpha\tau_j + (1-\alpha)\eta_j$ where η_j represents the binary heuristic information, τ_j is the quantity of pheromone accumulated at position j, and α is a parameter which determines the relative influence of the pheromone trail and the heuristic information.

Each ant always moves into one of the pixels of the nearest scanning line (line A, namely the shaded sets in figure 6), but the probability of transition combines the quality of each pixel in line A with that of a corresponding pixel in line B as indicated by the arrows in figure 6. Defining with l a pixel in line B corresponding to a pixel j in line A, the probability that ant k moves from position i to position j belonging to its feasible neighborhood N_i^k at step t is

$$p_{ij} = \frac{\frac{1}{d_j} \times [(1-\nu) \times q_j + \nu \times q_l]}{\sum\limits_{(j,l)\in N_j^k} \frac{1}{d_j} \times [(1-\nu) \times q_j + \nu \times q_l]} \qquad (1)$$

where ν is a parameter in a range $[0,1]$ indicating the ants

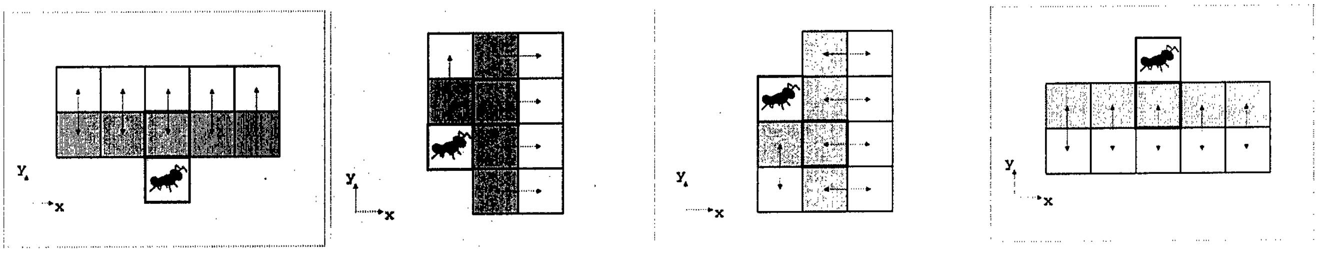


Fig. 6 Artificial ants move to one pixel of the shaded set (named *line A*) by calculating the quality of each pixel of line A and of the white region (named *line B*). The figures illustrate the set of pixels evaluated by ants when they cross region 1 (*a*), region 2 (*b*), region 3 (*c*), and region 4 (*d*).

field of view; for $\nu = 1$ the ant sees only line A pixels, for $\nu = 0$ the ant sees only line B pixels, while for intermediate values the ant focus of attention varies in between line A e B. d_j is the displacement of pixel j with respect to the central pixel of line A. The $1/d_j$ penalty favors straight trails in comparison with frequent small alternative leaps to the left and the right.

The system provides two different kinds of agents: purely stochastic ants and semi-deterministic ants. Both kinds choose their move with a uniformly distributed random rule, but the range of choice is different: purely stochastic ants have all the feasible neighborhood N illustrated in figure 6 at their disposal, while semi-deterministic ants choose only between the two pixels that have the highest p_{ij}. Both kinds of ants perform well on synthetic images; however, stochastic ants explore more widely the solution space but converge more slowly to a final solution than the semi-deterministic ants do. On the other hand, semi-deterministic ants follow well connected edges, but sometimes fail to find the best solution subspace in very irregular real images.

Once every ant has completed its tour, pheromone trails are updated through evaporation and reinforcement according to the following equation:

$$\tau_i(c+1) = (1-\rho) \times \tau_i(c) + \rho \times \left(\sum_{k=1}^{M} \Delta\tau_i^k + \Delta\tau_i^d \right) \qquad (2)$$

where ρ is the evaporation coefficient (ranging from 0 to 1), $\tau_i(c)$ is the quantity of pheromone present on pixel i at cycle c. Pheromone update $\Delta\tau$ is made up of two contributions. The first one is given by the sum of the pheromone deposed by each ant at the end of its tour. The second one is credited to the best trail according to an elitist strategy.

All ants are ranked according to the following rule: an ant obtains a high rank if it takes a long tour that passes through many edges or a low rank if it visits many pixels that are not edges. This rule is functional to the search of a good solution as it encourages ants to take the shortest path between two zones of connected pixels and does not pose any request on the total length of the trail.

The procedure described above is repeated for a number of cycles; experiments show that with 10 ants, 2 cycles are sufficient for a stable and reliable solution.

Finally, the output is the path of the ant of the highest rank in the last cycle.

A. *Preliminary results of shape detection*

Preliminary experiments were done mainly on synthetic images, like the one shown in figure 7.a. The performance of both purely stochastic and semi-deterministic ants were compared to deterministic edge trackers that follow the same movement rules as artificial ants, but proceed always on one deterministically chosen edge pixel. These trackers often failed to find a correct solution on both synthetic and natural images as edges often present bifurcations for which no deterministic decision rule could be conceived. The random decision rule of artificial ants, together with a high number of attempts, proved to perform better than deterministic trackers.

Only two cycles were sufficient to reach a feasible solution on synthetic images. Ten ants were running in each cycle. The evaporation coefficient ρ was set to 0 so that the second cycle would take into account the full information provided by the first. The second cycle proved useful in finding a stable solution, in the sense that the solution subspace detected by the system was kept the same over multiple attempts on the same image, in spite of the random nature of the algorithm.

Figure 7.d shows the path drawn by the best purely stochastic ant at the end of the second cycle on the normalized edge matrix of the synthetic pedestrian of figure 7.a. The ant correctly delineates the trunk and legs shape of the human shape, and, in this case, it cuts out the head. A future step in the recognition process will start from this shape and try to identify the head or possibly other parts of the body in their correct positions.

The result of figure 7.d was obtained using parameters $\alpha = 0.5, \nu = 0.2, \rho = 0.0, Q_a = 0.5, Q_d = 1.0, C = 2, M = 10$. Figure 7.b illustrates an image of the world-matrix on which artificial ants moved during the second cycle, where brighter pixels have a higher quality q_i. Figure 7.c shows the pheromone matrix τ_i. Again, the brighter the pixel, the higher its pheromone quantity.

IV. Discussion

This work presents a vision-based system for detecting pedestrians in road environments.

Initially, low-level processing techniques are used to focus on few areas of interest which potentially contain pedestrians. Then, a subsequent higher-level processing is used to further analyze these areas of the image by means of autonomous agents: an ant-based matching with a human shape model is used for validating the presence of pedestrians.

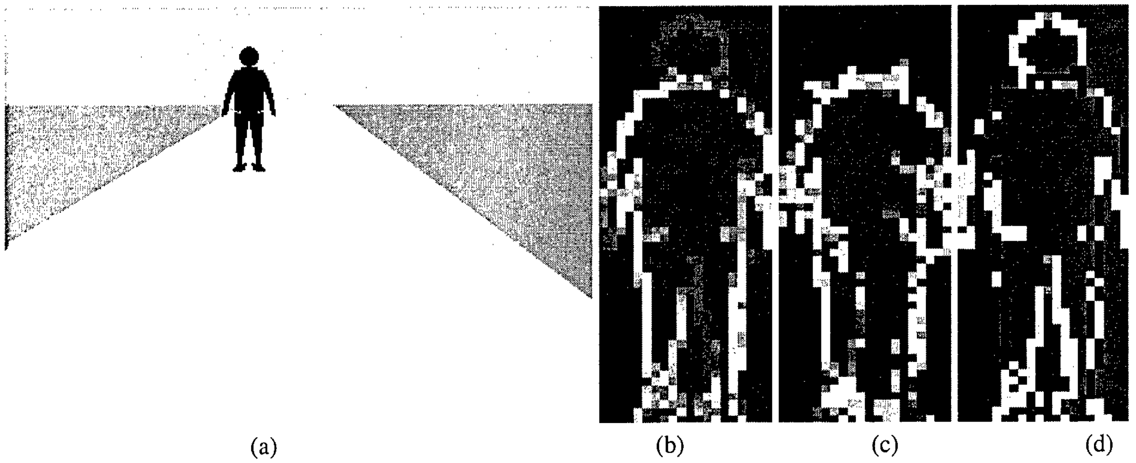


Fig. 7 (*a*) Example of a synthetic image used for experiments with artificial ants; (*b*) map of the world-matrix on which artificial ants move; (*c*) map of the pheromone trails deposed by ants after two cycles have been completed; (*d*) example of the result obtained by the best purely stochastic ant after two cycles.

This algorithm suits a medium distance search area. In fact, large bounding boxes may contain a too detailed shape, showing many disturbing small details that would certainly make their detection extremely difficult. In other words, the presence of texture (not only caused by different clothing) and the many different human postures that must be taken into account, would make the detection hard. On the other hand, very small bounding boxes enclosing far away pedestrians feature a very low information content. In these situations it is easy to obtain false positives, since many road participants (other than pedestrians), other objects, and even road infrastructures may present morphological characteristics similar to a human shape. It is therefore imperative to define a range of reasonable-sized bounding boxes in which the detection may lead to sufficiently accurate detections. In this work the considered size is: 12 x 28 pixel for the smallest bounding box, and 42 x 100 pixel for the largest one. This choice removes the small errors caused by false detections of small objects, as well as inaccurate detection (or even missed detections) of large pedestrians. Indeed this choice leads to a limited detection area in front of the vehicle. The system was tested on the images acquired by the vision system installed on-board of the ARGO experimental vehicle. With the current setup the search area ranges from 10 to 30 m.

The candidate selection procedure based on vertical symmetry and edge density proved to be a robust technique for focusing the attention on interesting regions. From the first preliminary results, the ant-based processing appears to be a promising method for detecting the contour of a human shape. To extend the detection to a larger set of pedestrian postures, other models are currently under development.

ACKNOWLEDGMENTS

The authors gratefully thank Dr. C. Adams and Dr. M. Del Rose from U. S. Army TACOM and Dr. S. Sampath from USARDSG-UK for their support in the research.

REFERENCES

[1] A. Broggi, M. Bertozzi, G. Conte, and A. Fascioli, "ARGO Prototype Vehicle," in *Intelligent Vehicle Technologies* (L. Vlacic, F. Harashima, and M. Parent, eds.), ch. 14, pp. 445–493, London, UK: Butterworth–Heinemann, June 2001. ISBN 0750650931.

[2] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking Pedestrian Recognition," in *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems'99*, (Tokyo, Japan), pp. 292–297, Oct. 1999.

[3] L. Zhao and C. Thorpe, "Stereo- and Neural Network-based Pedestrian Detection," in *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems'99*, (Tokyo, Japan), pp. 298–303, Oct. 1999.

[4] D. Beymer and K. Konolige, "Real-time Tracking of Multiple People using Continuous Detection," in *Procs. Intl. Conf. on Computer Vision*, 1999.

[5] D. M. Gavrila, "Pedestrian Detection from a Moving Vehicle," in *Procs. of European Conf. on Computer Vision*, vol. 2, pp. 37–49, June–July 2000.

[6] D. M. Gavrila, "Sensor-based Pedestrian Protection," *IEEE Intelligent Systems*, vol. 16, pp. 77–81, Nov.–Dec. 2001.

[7] C. Papageorgiou, T. Evgeniou, and T. Poggio, "A Trainable Pedestrian Detection System," in *Procs. IEEE Intelligent Vehicles Symposium'98*, (Stuttgart, Germany), pp. 241–246, Oct. 1998.

[8] R. Polana and R. C. Nelson, "Detection and Recognition of Periodic, Nonrigid Motion," *Internation Journal of Computer Vision*, vol. 23, pp. 261–282, June–July 1997.

[9] S. J. McKenna and S. Gong, "Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition," *Lecture Notes in Computer Science*, vol. 1206, pp. 177–184, Mar. 1997.

[10] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 781–796, Aug. 2000.

[11] C. Wöhler, J. K. Aulaf, T. Pörtner, and U. Franke, "A Time Delay Neural Network Algorithm for Real-time Pedestrian Detection," in *Procs. IEEE Intelligent Vehicles Symp.'98*, (Germany), pp. 247–251, Oct. 1998.

[12] C. Wöhler, U. Kreßel, and J. K. Anlauf, "Pedestrian Recognition by Classification of Image Sequences – Global Approaches vs. Local Spatio-Temporal Processing," in *Procs. IEEE Intl. Conf. on Pattern Recognition*, (Barcelona, Spain), Sept. 2000.

[13] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking Pedestrian Recognition," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 155–163, Sept. 2000.

[14] V. Philomin, R. Duraiswami, and L. Davis, "Pedestrian Tracking from a Moving Vehicle," in *Procs. IEEE Intelligent Vehicles Symposium 2000*, (Detroit, USA), pp. 350–355, Oct. 2000.

[15] L. Zhao and C. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 148–154, Sept. 2000.

[16] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context Free Attentional Operators: the Generalized Symmetry Transform," *Intl. Journal of Computer Vision, Special Issue on Qualitative Vision*, vol. 14, pp. 119–130, 1994.

[17] M. Dorigo and G. Di Caro, "The ant colony optimization meta-heuristic," in *New Ideas in Optimization* (D. Corne, M. Dorigo, and F. Glover, eds.), pp. 11–32, London, UK: McGraw-Hill, 1999.

[18] M. Dorigo and L. M. Gambardella, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem," *IEEE Tran. on Evolutionary Computation*, vol. 1, pp. 53–66, Apr. 1997.

Vision-based Pedestrian Detection: will Ants Help?

M. Bertozzi, A. Broggi, A. Fascioli, and P. Lombardi

Abstract— **This work presents the vision-based system for detecting pedestrians in road environments implemented on the ARGO prototype vehicle developed by the University of Parma. The system is aimed at the localization of pedestrians in various poses, positions and clothing, and is not limited to moving people.**

Initially, attentive vision techniques relying on the search for specific characteristics of pedestrians such as vertical symmetry and strong presence of edges, allow to select interesting regions likely to contain pedestrians. Then, such candidates areas are validated verifying the actual presence of pedestrians by means of an shape detection technique based on the application of autonomous agents.

Keywords— **pedestrian detection, machine vision, autonomous agents, symmetry, ants.**

I. Introduction

The detection of pedestrians is a mandatory requirement for future driving assistance systems, since having the capability of avoiding crushes with pedestrians is essential for effectively aiding the driver in urban environments.

This work presents the vision-based system for detecting pedestrians implemented on the ARGO vehicle. ARGO is an experimental autonomous vehicle developed by the University of Parma, equipped with a vision system, a processing engine, and automatic steering capabilities [1]. The main target of the ARGO Project is the development of an active safety system which can also act as an automatic pilot for a standard road vehicle.

Vision-based pedestrian detection in outdoor scenes is a challenging task even in the case of a stationary camera. In fact, pedestrians usually wear different clothes with various colors that, sometimes, are barely distinguishable from the background (this is particularly true when processing black and white images). Moreover, pedestrians can wear or carry items like hats, bags, umbrellas, and many others, which give a broad variability to their shape.

When the vision system is installed on-board of a moving vehicle additional problems must be faced, since the observer's ego-motion entails additional motion in the background and changes in the illumination conditions. In addition, since Pedestrian Detection is more likely to be of use in a urban environment, also the presence of a complex background (including buildings, moving or parked cars, cycles, road signs, signals, ...) must be taken into account.

Widely used approaches for addressing vision-based Pedestrian Detection are: the search of specific patterns or textures [2], stereo vision [3, 4], shape detection [5, 6, 7], motion detection [8, 9, 10], neural networks [11, 12]. The great part of the research groups use a combination of two or more of these approaches [13, 14, 15]. Anyway, only a few of these systems have already proved their efficacy in applications for intelligent vehicles.

In this work the strong vertical symmetry of the human shape is exploited to determine specific regions of interest which are likely to contain pedestrians. Subsequently, an agent-based algorithm is used to verify the presence of pedestrians in those areas through an innovative application of a shape detection technique. This method allows the identification of pedestrians in various poses, positions and clothing, and is not limited to walking people.

This paper is organized as follows. Section 2 introduces the attentive algorithm aimed at the determination of the areas of interests, and shows its results. Section 3 describes the agent-based algorithm used for the validation of the candidates and presents the preliminary results of this shape detection technique. Section 4 ends the paper with some final remarks.

II. Attentive vision

As a first processing step, attentive vision techniques are applied to concentrate the analysis on specific regions of interest only. In fact, the aim of the low-level part of the processing is the focusing on potential candidate areas to be further examined at a higher-level stage in the following steps.

The areas considered as candidate are rectangular bounding boxes which:

- have a size in pixels deriving from the knowledge of the intrinsic parameters of the vision system (angular aperture and resolution); in other words, once defined the size and distance of a pedestrian in the 3D world (e. g. 1.8 $m \times 0.6$ m at 20 m), simple perspective considerations give the size in pixel of its projection in the image (see figure 1.a);
- enclose a portion of the image which exhibits the low-level features that characterize the presence of a pedestrian, i. e. a strong vertical symmetry and a high density of vertical edges.

These bounding boxes will be then checked against a human shape model, taking into account the contour of the object they enclose, in order to be validated.

The search for candidates would require an exhaustive search in the whole image (see figure 1.b). However, the knowledge of the system's extrinsic parameters, together with a flat scene assumption (see figure 1.c), is exploited to limit the analysis to a stripe of the image (hereinafter referred to as *search area*). The displacement of this stripe depends on the pedestrian's distance, while its height is related to the pedestrian's height (see figure 1.d). Since by definition a pedestrian is a human shaped road participant, the flat world assumption becomes an assumption on the road slope, which is anyway a loose hypothesis in a road environment, particularly in the area immediately ahead of the vehicle. Besides the obvious advantage of avoiding false detections in wrong areas, the processing of the search

This work was supported by the European Research Office of the U. S. Army under contract number N68171-01-M-5857

M. Bertozzi, A. Broggi, and A. Fascioli are with the Dip. di Ingegneria dell'Informazione, Università di Parma, ITALY. E-mail: {bertozzi,broggi,fascioli}@ce.unipr.it.

P. Lombardi is with the Dip. di Informatica e Sistemistica, Università di Pavia, ITALY. E-mail: lombardi@vision.unipv.it

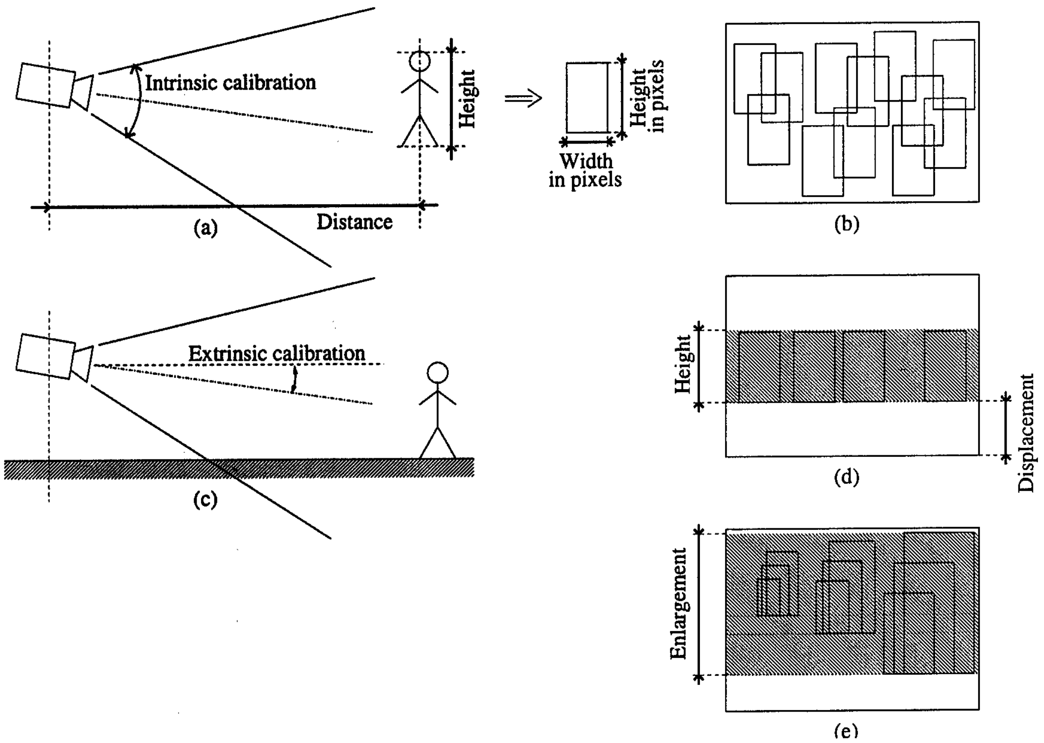


Fig. 1 *(a)* Computation of the bounding box size given the intrinsic parameters and the size and distance of a pedestrian; *(b)* exhaustive search for candidates in the whole image; *(c)* the search area can be limited to a stripe given the extrinsic parameters and a flat scene assumption; *(d)* the displacement and height of the stripe depend on the pedestrian distance and height, respectively; *(e)* the search area is enlarged to explore a range of distances and heights.

area only reduces the computational time. Indeed, the analysis cannot be limited to a fixed size and distance of the target and a given range for each parameter is in fact explored (e. g. $1.6 \div 2.0\ m \times 0.5 \div 0.7\ m$ at $10 \div 30\ m$). The introduction of these ranges generates two further degrees of freedom in the size and position of the bounding boxes. In other words, the search area is enlarged to accommodate all possible combinations of height, width, and distance (see figure 1.e).

The analysis proceeds in this way: the columns of the image are considered as possible symmetry axes for bounding boxes. For each symmetry axis different bounding boxes are evaluated scanning a specific range of distances from the camera (the distance determines the position of the bounding box base) and a reasonable range of heights and widths for a pedestrian (the corresponding bounding box size can be computed through the calibration).

However, not all the possible symmetry axes are considered: since edges are chosen as discriminant in most of the following analysis, a pre-attentive filter is applied, aimed at the selection of the areas with a high density of edges. In particular, for each axis the count of edge pixel is computed in a portion of the search area centered on the axis itself and as wide as the maximum bounding box width. Axes centered on regions which contain a number of edges lower than the average value are then dropped.

For each of the remaining axes the best candidate area is selected among the bounding boxes which share that symmetry axis, while having different position (base) and size (height and width). Vertical symmetry has been chosen as a main distinctive feature for pedestrians. Symmetry edge maps, e. g. the Generalized Symmetry Transform (GST) [16], have already been proposed as methods to locate interest points in the image prior to any segmentation or extraction of context-dependent information. Unfortunately, these methods are generally computationally expensive. Alternatively, two different symmetry measures are performed: one on the gray-level values (G) and one on the gradient values, considering only edges with a vertical direction (E). The selection of the best bounding box is based on maximizing a linear combination of the two symmetry measures, masked by the density of edges in the box (D), as shown in the following equation: $S = (k1 \times G + k2 \times E) \times D$. The weights $k1$ and $k2$ were experimentally determined analyzing a large number of images. Figure 2 shows the original input image, the result of a clustering operation used to improve the detection of edges, a binary image containing the vertical edges, and a number of histograms representing the maximum (i) symmetry of gray-levels, (ii) symmetry of vertical edges, and (iii) density of vertical edges among the bounding boxes examined for each axis. The histogram in figure 2.g represents the linear combination of all the above. The histograms are actually computed only for the regions of the search area with a high density of edges, however in figure 2 they are completely displayed for a better understanding. It is evident that, using the density of vertical

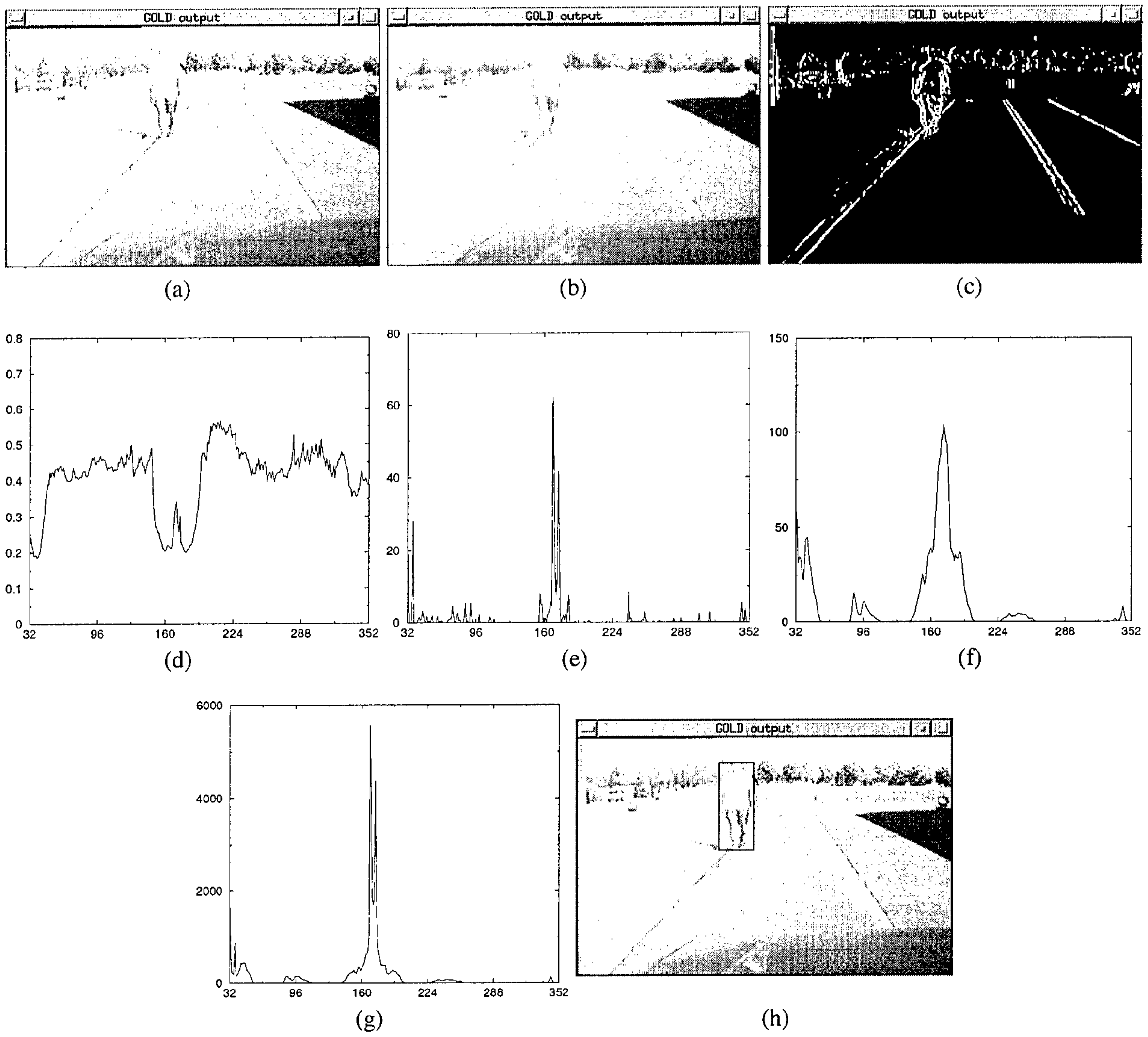


Fig. 2 Intermediate results leading to the localization of bounding boxes: *(a)* original image; *(b)* clusterized image; *(c)* vertical edges; *(d)* histogram representing grey level symmetries; *(e)* histogram representing vertical edges symmetries; *(f)* histogram representing vertical edges density; *(g)* histogram representing the overall symmetry S for the best bounding box for each column; *(h)* the resulting bounding box.

edges as a mask, interesting areas present high values for both the symmetry of gray-levels and symmetry of vertical edges. The resulting histogram is therefore thresholded and its over-threshold peaks are selected as representing candidate bounding boxes.

An adjustment of the bounding boxes' size is yet needed. In fact, when comparing the gray-level symmetry of different bounding boxes centered on the same axis, larger boxes tend to overcome smaller ones since pedestrians are generally surrounded by homogeneous areas such as concrete underneath or the sky above (this is true for other objects, too). Therefore, the bounding box which presents the maximum symmetry tends to be larger than the object it contains because it includes uniform regions. For this reason, given a peak of the overall histogram representing a selected symmetry axis, the exact height and width of the best bounding box are actually taken as those possessed by the box which maximizes a new function among the ones having the same axis. This function is computed as the product of the symmetry of vertical edges (E) and density of vertical edges (D) only. Figure 3 summarizes the overall candidate generation process.

The result of this low-level processing is a list of candidate bounding boxes which is fed to the following stage, whose task is their validation as pedestrians, based on higher-level characteristics.

A. *Results of low-level attentive vision*

The algorithm has been tested on a large number of images acquired in different situations ranging from simple uncluttered scenes to complex scenarios. As an example, figure 4 shows the result of the selection of candidate bounding boxes in three different situations. In figure 4.a a correct detection of two pedes-

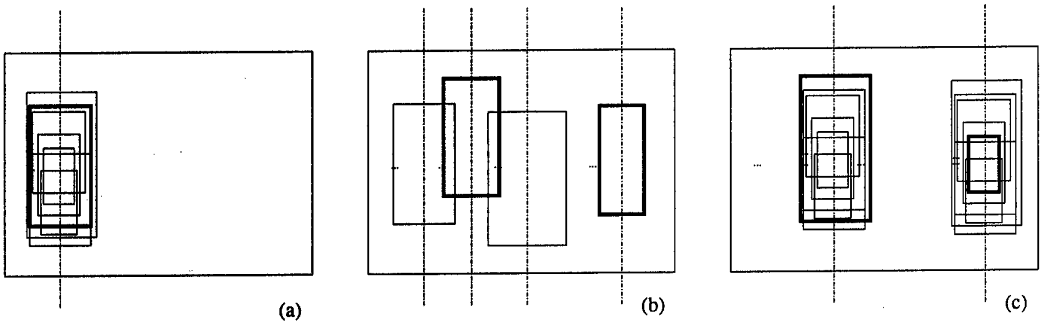


Fig. 3 *(a)* Selection of the best bounding box for each symmetry axis; *(b)* selection of the best symmetry axes; *(c)* selection of the best candidates for each selected axis by choosing the bounding box which maximizes the symmetry and density of vertical edges.

trians is displayed. Figure 4.b presents a complex scenario in which only the central pedestrian is detected; two other pedestrians are missed because (i) the first is confused with the background, and (ii) the second is only partially visible; moreover, the high symmetry of a tree has been detected. In figure 4.c the two crossing pedestrians have been localized, but other symmetrical areas are highlighted as well; the first is due to the symmetry of the area between two pedestrians, and the other is due to the presence of symmetrical road infrastructures between two trees.

Some general considerations can be drawn on the behavior of this candidate selection procedure. In situations in which pedestrians are sufficiently contrasted with respect to the background and completely visible (i. e. not hidden by other pedestrians or objects) the localization of pedestrians based on symmetry and edge density proves to be robust. Thanks to the use of vertical edges the width of the bounding boxes enclosing pedestrians is generally determined with a good precision. On the other hand, a lower accuracy is obtained for the localization of the top and bottom of the bounding box. A refinement of the bounding box height is under development.

Symmetrical objects other than pedestrians may happen to be detected as well. In order to get rid of such false positives a number of filters have been devised which rely on the analysis of the distribution of edges within the bounding box. These filters, which are still under evaluation, show promising results regarding the elimination of both artifacts (such as poles, road signs, buildings, and other road infrastructures) and symmetrical areas given by a uniform portion of the background between two foreground objects with similar lateral borders (see figure 4.c).

III. SHAPE DETECTION USING AUTONOMOUS AGENTS

This section describes the shape detection technique.

Different edges are selected and connected, where possible, in order to form a contour. Thanks to the way the contour is built, it will represent the shape of the pedestrian body. Matching techniques may be used in regions of the bounding box that lie in the correct position in relation to the formulated hypothesis of the pedestrian.

Essentially, the process consists in adapting a deformable coarse model to the bounding boxes. Thanks to its roughness the model is sufficiently general and can be adapted to a variety of postures. Anyway, it is limited to standing pedestrians. The model adjustment is done through an evolutionary approach with a number of independent agents acting as edge trackers. The agents explore a feature map displaying the edges contained in a given bounding box and stochastically build hypotheses of a feasible contour of a human body. The idea is taken from the Ant Colony Optimization (ACO) metaheuristic devised to solve hard combinatorial optimization problems, originally inspired by the communication behavior of real ants [17]. The system proposed here is a transposition to image analysis of one of the first ACO algorithms, the AS-*cycle*.

In nature, when ants look for food, they communicate the path and the outcome of their exploration to other ants by marking their path with a pheromone trail, its intensity depending on the distance of the food from the nest, and on its quality and quantity. Other ants are attracted by strong pheromone trails, thus the path to an abundant food source close to the nest is marked again and again until it becomes more frequented and even more attractive.

This concept can be applied to the analysis of an image by creating a colony of artificial ants that looks for an optimal combination of edge pixels that maximizes the coherency of their position according to a given model. Each ant in turn traces a solution in a solution space made up of all the possible paths connecting two pixels in a matrix. The decisional basis for each step of an ant is provided by two factors: one is a local heuristic η_i that quantifies the attractiveness of pixel i for its intrinsic characteristics; the second is the information on that pixel made available by previous attempts of other ants, in the form of a quantity of pheromone τ_i.

Artificial ants explore a world which is a matrix of pixels derived by the resampling of the edge map of the bounding box under analysis. In our experiments, the normalized world-matrix is sized 20 × 45 pixels. Each pixel i is initialized with a binary value: 1 if it contains an edge, 0 if not. This value represents its intrinsic attractiveness η_i and is the basis for the heuristic research. All the pheromone τ_i is initialized at 0. The world-matrix is visited by M ants in parallel, and the process is repeated for C cycles. At the end of each cycle, new pheromone is deposed on the trails pursued by the ants, and some of that

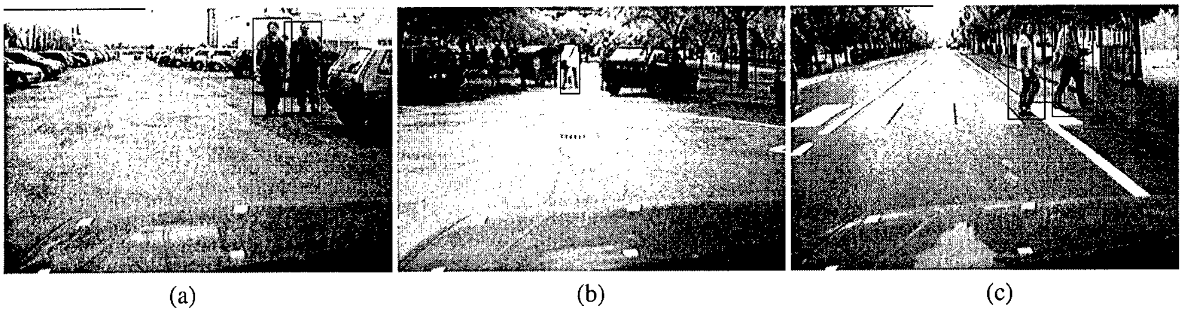


Fig. 4 Result of low-level processing in different situations: *(a)* a correct detection of two pedestrians *(b)* a complex scenario in which only the central pedestrian is detected, and the high symmetry of a tree has been detected as well; *(c)* two crossing pedestrians have been localized, but other symmetrical areas are highlighted as well.

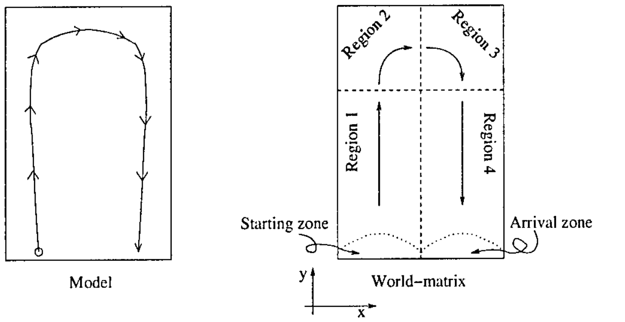


Fig. 5 Artificial ants move through the world-matrix starting from the left half of the lower border, and moving through regions 1, 2, 3 and 4 until they reach the arrival line.

accumulated evaporates. In this way, solutions built several cycles before, progressively loose their importance. On the other hand, pheromone on pixels that compose the path of frequently selected solutions grows. and eventually this information surpasses that given by the heuristic.

A crucial point to be understood is that artificial ants do not need to reach an optimal solution in the edge connection problem. Often, no real optimal solution exists even to a human inspector. The colony needs only to find a sufficiently valuable path that permits to continue the recognition, free of noisy edge pixels.

The ant system develops from a very elastic and deformable coarse model of a human body. The model is encoded in the progression rules that guide the ants through the solution space. The rules effectively restrict the whole space of possible solutions to a subspace that includes the searched shape. The system will then provide attempts to find a feasible path in this subspace, and each attempt will be evaluated by a confidence function as it is detailed in the following.

All ants start from the left half of the lower side of the world-matrix. The world is divided into four regions, as detailed in figure 5. In each region ants proceed of one step forward in the direction of one of the axis (y in regions 1 and 4, and x in regions 2 and 3), and can choose among a set of s pixels lying on the line or column in front of them. Additionally, in regions 2 and 3 ants have the option of moving vertically, thus they can follow very steep edges as well as very flat ones.

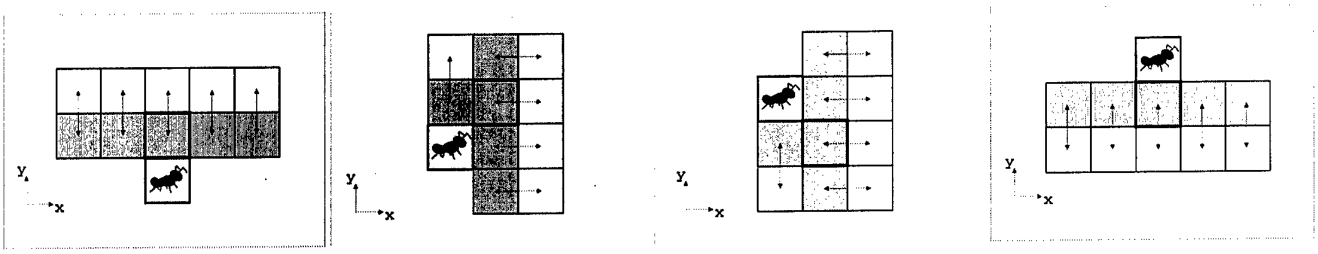
The starting point of each ant is chosen randomly among the edge pixels lying in the starting region, i.e. the left half of the lower border. If no edge pixel is present, the starting point is set on a random point belonging to the region. The choice of starting from edge pixels does not pose a hard restriction to the exploration of the solution space: the bounding box usually comprises edges in the lowest line owing to the mechanism that determines its dimensions based on the edge density. Most of the times, the edges appearing on the lower line of a well-centered bounding box correspond to the feet of the pedestrian. Each ant stops its journey when it reaches the right half of the lower border of the world-matrix.

An ant is an independent pixel-sized agent; it has a local exploratory capability, limited to the set of pixels belonging to the scanning region N as described above, and of those lying on the following line as well. Figure 6 illustrates the situation for the scanning sets for each region of the world-matrix. Each pixel under consideration is associated to a quality measure that takes into account terms pertaining to both the feature map of the edges, and the pheromone deposed by previous ants. The quality of pixel j is expressed as $q_j = \alpha\tau_j + (1-\alpha)\eta_j$ where η_j represents the binary heuristic information, τ_j is the quantity of pheromone accumulated at position j, and α is a parameter which determines the relative influence of the pheromone trail and the heuristic information.

Each ant always moves into one of the pixels of the nearest scanning line (line A, namely the shaded sets in figure 6), but the probability of transition combines the quality of each pixel in line A with that of a corresponding pixel in line B as indicated by the arrows in figure 6. Defining with l a pixel in line B corresponding to a pixel j in line A, the probability that ant k moves from position i to position j belonging to its feasible neighborhood N_i^k at step t is

$$p_{ij} = \frac{\frac{1}{d_j} \times [(1-\nu) \times q_j + \nu \times q_l]}{\sum\limits_{(j,l)\in N_j^k} \frac{1}{d_j} \times [(1-\nu) \times q_j + \nu \times q_l]} \qquad (1)$$

where ν is a parameter in a range $[0,1]$ indicating the ants



(a) (b) (c) (d)

Fig. 6 Artificial ants move to one pixel of the shaded set (named *line A*) by calculating the quality of each pixel of line A and of the white region (named *line B*). The figures illustrate the set of pixels evaluated by ants when they cross region 1 (*a*), region 2 (*b*), region 3 (*c*), and region 4 (*d*).

field of view; for $\nu = 1$ the ant sees only line A pixels, for $\nu = 0$ the ant sees only line B pixels, while for intermediate values the ant focus of attention varies in between line A e B. d_j is the displacement of pixel j with respect to the central pixel of line A. The $1/d_j$ penalty favors straight trails in comparison with frequent small alternative leaps to the left and the right.

The system provides two different kinds of agents: purely stochastic ants and semi-deterministic ants. Both kinds choose their move with a uniformly distributed random rule, but the range of choice is different: purely stochastic ants have all the feasible neighborhood N illustrated in figure 6 at their disposal, while semi-deterministic ants choose only between the two pixels that have the highest p_{ij}. Both kinds of ants perform well on synthetic images; however, stochastic ants explore more widely the solution space but converge more slowly to a final solution than the semi-deterministic ants do. On the other hand, semi-deterministic ants follow well connected edges, but sometimes fail to find the best solution subspace in very irregular real images.

Once every ant has completed its tour, pheromone trails are updated through evaporation and reinforcement according to the following equation:

$$\tau_i(c+1) = (1-\rho)\times\tau_i(c) + \rho\times\left(\sum_{k=1}^{M}\Delta\tau_i^k + \Delta\tau_i^d\right) \tag{2}$$

where ρ is the evaporation coefficient (ranging from 0 to 1), $\tau_i(c)$ is the quantity of pheromone present on pixel i at cycle c. Pheromone update $\Delta\tau$ is made up of two contributions. The first one is given by the sum of the pheromone deposed by each ant at the end of its tour. The second one is credited to the best trail according to an elitist strategy.

All ants are ranked according to the following rule: an ant obtains a high rank if it takes a long tour that passes through many edges or a low rank if it visits many pixels that are not edges. This rule is functional to the search of a good solution as it encourages ants to take the shortest path between two zones of connected pixels and does not pose any request on the total length of the trail.

The procedure described above is repeated for a number of cycles; experiments show that with 10 ants, 2 cycles are sufficient for a stable and reliable solution.

Finally, the output is the path of the ant of the highest rank in the last cycle.

A. Preliminary results of shape detection

Preliminary experiments were done mainly on synthetic images, like the one shown in figure 7.a. The performance of both purely stochastic and semi-deterministic ants were compared to deterministic edge trackers that follow the same movement rules as artificial ants, but proceed always on one deterministically chosen edge pixel. These trackers often failed to find a correct solution on both synthetic and natural images as edges often present bifurcations for which no deterministic decision rule could be conceived. The random decision rule of artificial ants, together with a high number of attempts, proved to perform better than deterministic trackers.

Only two cycles were sufficient to reach a feasible solution on synthetic images. Ten ants were running in each cycle. The evaporation coefficient ρ was set to 0 so that the second cycle would take into account the full information provided by the first. The second cycle proved useful in finding a stable solution, in the sense that the solution subspace detected by the system was kept the same over multiple attempts on the same image, in spite of the random nature of the algorithm.

Figure 7.d shows the path drawn by the best purely stochastic ant at the end of the second cycle on the normalized edge matrix of the synthetic pedestrian of figure 7.a. The ant correctly delineates the trunk and legs shape of the human shape, and, in this case, it cuts out the head. A future step in the recognition process will start from this shape and try to identify the head or possibly other parts of the body in their correct positions.

The result of figure 7.d was obtained using parameters $\alpha = 0.5, \nu = 0.2, \rho = 0.0, Q_a = 0.5, Q_d = 1.0, C = 2, M = 10$. Figure 7.b illustrates an image of the world-matrix on which artificial ants moved during the second cycle, where brighter pixels have a higher quality q_i. Figure 7.c shows the pheromone matrix τ_i. Again, the brighter the pixel, the higher its pheromone quantity.

IV. Discussion

This work presents a vision-based system for detecting pedestrians in road environments.

Initially, low-level processing techniques are used to focus on few areas of interest which potentially contain pedestrians. Then, a subsequent higher-level processing is used to further analyze these areas of the image by means of autonomous agents: an ant-based matching with a human shape model is used for validating the presence of pedestrians.

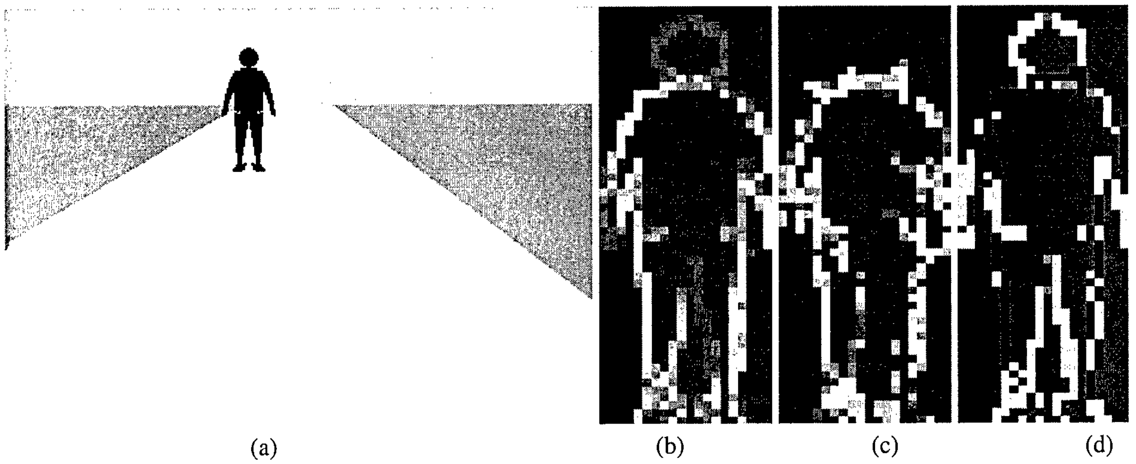


Fig. 7 (*a*) Example of a synthetic image used for experiments with artificial ants; (*b*) map of the world-matrix on which artificial ants move; (*c*) map of the pheromone trails deposed by ants after two cycles have been completed; (*d*) example of the result obtained by the best purely stochastic ant after two cycles.

This algorithm suits a medium distance search area. In fact, large bounding boxes may contain a too detailed shape, showing many disturbing small details that would certainly make their detection extremely difficult. In other words, the presence of texture (not only caused by different clothing) and the many different human postures that must be taken into account, would make the detection hard. On the other hand, very small bounding boxes enclosing far away pedestrians feature a very low information content. In these situations it is easy to obtain false positives, since many road participants (other than pedestrians), other objects, and even road infrastructures may present morphological characteristics similar to a human shape. It is therefore imperative to define a range of reasonable-sized bounding boxes in which the detection may lead to sufficiently accurate detections. In this work the considered size is: 12 x 28 pixel for the smallest bounding box, and 42 x 100 pixel for the largest one. This choice removes the small errors caused by false detections of small objects, as well as inaccurate detection (or even missed detections) of large pedestrians. Indeed this choice leads to a limited detection area in front of the vehicle. The system was tested on the images acquired by the vision system installed on-board of the ARGO experimental vehicle. With the current setup the search area ranges from 10 to 30 m.

The candidate selection procedure based on vertical symmetry and edge density proved to be a robust technique for focusing the attention on interesting regions. From the first preliminary results, the ant-based processing appears to be a promising method for detecting the contour of a human shape. To extend the detection to a larger set of pedestrian postures, other models are currently under development.

ACKNOWLEDGMENTS

The authors gratefully thank Dr. C. Adams and Dr. M. Del Rose from U. S. Army TACOM and Dr. S. Sampath from USARDSG-UK for their support in the research.

REFERENCES

[1] A. Broggi, M. Bertozzi, G. Conte, and A. Fascioli, "ARGO Prototype Vehicle," in *Intelligent Vehicle Technologies* (L. Vlacic, F. Harashima, and M. Parent, eds.), ch. 14, pp. 445–493, London, UK: Butterworth–Heinemann, June 2001. ISBN 0750650931.

[2] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking Pedestrian Recognition," in *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems'99*, (Tokyo, Japan), pp. 292–297, Oct. 1999.

[3] L. Zhao and C. Thorpe, "Stereo- and Neural Network-based Pedestrian Detection," in *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems'99*, (Tokyo, Japan), pp. 298–303, Oct. 1999.

[4] D. Beymer and K. Konolige, "Real-time Tracking of Multiple People using Continuous Detection," in *Procs. Intl. Conf. on Computer Vision*, 1999.

[5] D. M. Gavrila, "Pedestrian Detection from a Moving Vehicle," in *Procs. of European Conf. on Computer Vision*, vol. 2, pp. 37–49, June–July 2000.

[6] D. M. Gavrila, "Sensor-based Pedestrian Protection," *IEEE Intelligent Systems*, vol. 16, pp. 77–81, Nov.–Dec. 2001.

[7] C. Papageorgiou, T. Evgeniou, and T. Poggio, "A Trainable Pedestrian Detection System," in *Procs. IEEE Intelligent Vehicles Symposium'98*, (Stuttgart, Germany), pp. 241–246, Oct. 1998.

[8] R. Polana and R. C. Nelson, "Detection and Recognition of Periodic, Nonrigid Motion," *Internation Journal of Computer Vision*, vol. 23, pp. 261–282, June–July 1997.

[9] S. J. McKenna and S. Gong, "Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition," *Lecture Notes in Computer Science*, vol. 1206, pp. 177–184, Mar. 1997.

[10] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 781–796, Aug. 2000.

[11] C. Wöhler, J. K. Aulaf, T. Pörtner, and U. Franke, "A Time Delay Neural Network Algorithm for Real-time Pedestrian Detection," in *Procs. IEEE Intelligent Vehicles Symp. '98*, (Germany), pp. 247–251, Oct. 1998.

[12] C. Wöhler, U. Kreßel, and J. K. Anlauf, "Pedestrian Recognition by Classification of Image Sequences – Global Approaches vs. Local Spatio-Temporal Processing," in *Procs. IEEE Intl. Conf. on Pattern Recognition*, (Barcelona, Spain), Sept. 2000.

[13] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking Pedestrian Recognition," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 155–163, Sept. 2000.

[14] V. Philomin, R. Duraiswami, and L. Davis, "Pedestrian Tracking from a Moving Vehicle," in *Procs. IEEE Intelligent Vehicles Symposium 2000*, (Detroit, USA), pp. 350–355, Oct. 2000.

[15] L. Zhao and C. Thorpe, "Stereo and neural network-based pedestrian detection," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, pp. 148–154, Sept. 2000.

[16] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context Free Attentional Operators: the Generalized Symmetry Transform," *Intl. Journal of Computer Vision, Special Issue on Qualitative Vision*, vol. 14, pp. 119–130, 1994.

[17] M. Dorigo and G. Di Caro, "The ant colony optimization meta-heuristic," in *New Ideas in Optimization* (D. Corne, M. Dorigo, and F. Glover, eds.), pp. 11–32, London, UK: McGraw-Hill, 1999.

[18] M. Dorigo and L. M. Gambardella, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem," *IEEE Tran. on Evolutionary Computation*, vol. 1, pp. 53–66, Apr. 1997.